

# Big Data's Disparate Impact

Solon Barocas\* & Andrew D. Selbst\*\*

*Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers. In other cases, data may simply reflect the widespread biases that persist in society at large. In still others, data mining can discover surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality. Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.*

*This Essay examines these concerns through the lens of American antidiscrimination law—more particularly, through Title*

---

DOI: <http://dx.doi.org/10.15779/Z38BG31>

California Law Review, Inc. (CLR) is a California nonprofit corporation. CLR and the authors are solely responsible for the content of their publications.

\* Postdoctoral Research Associate, Center for Information Technology Policy, Princeton University; Ph.D. 2014, New York University, Department of Media, Culture, and Communication. This research was supported in part by the Center for Information Technology Policy at Princeton University.

\*\* Scholar in Residence, Electronic Privacy Information Center; Visiting Researcher, Georgetown University Law Center; Visiting Fellow, Yale Information Society Project; J.D. 2011, University of Michigan Law School. The authors would like to thank Jane Bambauer, Alvaro Bedoya, Marjory Blumenthal, Danielle Citron, James Grimmelman, Moritz Hardt, Don Herzog, Janine Hiller, Chris Hoofnagle, Joanna Huey, Patrick Ishizuka, Michael Kirkpatrick, Aaron Konopasky, Joshua Kroll, Mark MacCarthy, Arvind Narayanan, Helen Norton, Paul Ohm, Scott Peppet, Joel Reidenberg, David Robinson, Kathy Strandburg, David Vladeck, members of the Privacy Research Group at New York University, and the participants of the 2014 Privacy Law Scholars Conference for their helpful comments. Special thanks also to Helen Nissenbaum and the Information Law Institute at New York University for giving us an interdisciplinary space to share ideas, allowing this paper to come about. Copyright © 2016 by Solon Barocas and Andrew Selbst. This Essay is available for reuse under the Creative Commons Attribution-ShareAlike 4.0 International License, <http://creativecommons.org/licenses/by-sa/4.0/>. The required attribution notice under the license must include the article's full citation information, e.g., "Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016)."

*VII's prohibition of discrimination in employment. In the absence of a demonstrable intent to discriminate, the best doctrinal hope for data mining's victims would seem to lie in disparate impact doctrine. Case law and the Equal Employment Opportunity Commission's Uniform Guidelines, though, hold that a practice can be justified as a business necessity when its outcomes are predictive of future employment outcomes, and data mining is specifically designed to find such statistical correlations. Unless there is a reasonably practical way to demonstrate that these discoveries are spurious, Title VII would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others' discrimination against members of protected groups, or flaws in the underlying data.*

*Addressing the sources of this unintentional discrimination and remedying the corresponding deficiencies in the law will be difficult technically, difficult legally, and difficult politically. There are a number of practical limits to what can be accomplished computationally. For example, when discrimination occurs because the data being mined is itself a result of past intentional discrimination, there is frequently no obvious method to adjust historical data to rid it of this taint. Corrective measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain. These challenges for reform throw into stark relief the tension between the two major theories underlying antidiscrimination law: anticlassification and antisubordination. Finding a solution to big data's disparate impact will require more than best efforts to stamp out prejudice and bias; it will require a wholesale reexamination of the meanings of "discrimination" and "fairness."*

Introduction .....	673
I. How Data Mining Discriminates.....	677
A. Defining the "Target Variable" and "Class Labels" .....	677
B. Training Data .....	680
1. Labeling Examples .....	681
2. Data Collection .....	684
C. Feature Selection .....	688
D. Proxies .....	691
E. Masking .....	692
II. Title VII Liability for Discriminatory Data Mining.....	694
A. Disparate Treatment.....	694
B. Disparate Impact.....	701
C. Masking and Problems of Proof .....	712
III. The Difficulty for Reforms .....	714
A. Internal Difficulties.....	715

1. Defining the Target Variable .....	715
2. Training Data .....	716
a. Labeling Examples.....	716
b. Data Collection .....	717
3. Feature Selection .....	719
4. Proxies .....	720
B. External Difficulties.....	723
Conclusion .....	729

## INTRODUCTION

“Big Data” is the buzzword of the decade.<sup>1</sup> Advertisers want data to reach profitable consumers,<sup>2</sup> medical professionals to find side effects of prescription drugs,<sup>3</sup> supply-chain operators to optimize their delivery routes,<sup>4</sup> police to determine where to focus resources,<sup>5</sup> and social scientists to study human interactions.<sup>6</sup> Though useful, however, data is not a panacea. Where data is used predictively to assist decision making, it can affect the fortunes of whole classes of people in consistently unfavorable ways. Sorting and selecting for the best or most profitable candidates means generating a model with winners and losers. If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.

Although we live in the post–civil rights era, discrimination persists in American society and is stubbornly pervasive in employment, housing, credit, and consumer markets.<sup>7</sup> While discrimination certainly endures in part due to decision makers’ prejudices, a great deal of modern-day inequality can be attributed to what sociologists call “institutional” discrimination.<sup>8</sup> Unconscious, implicit biases and inertia within society’s institutions, rather than intentional

---

1. *Contra* Sanjeev Sardana, *Big Data: It’s Not a Buzzword, It’s a Movement*, FORBES (Nov. 20, 2013), <http://www.forbes.com/sites/sanjeevsardana/2013/11/20/bigdata> [https://perma.cc/9Y37-ZFT5].

2. Tanzina Vega, *New Ways Marketers Are Manipulating Data to Influence You*, N.Y. TIMES: BITS (June 19, 2013, 9:49 PM), <http://bits.blogs.nytimes.com/2013/06/19/new-ways-marketers-are-manipulating-data-to-influence-you> [https://perma.cc/238F-9T8X].

3. Nell Greenfieldboyce, *Big Data Peeps at Your Medical Records to Find Drug Problems*, NPR (July 21, 2014, 5:15 AM), <http://www.npr.org/blogs/health/2014/07/21/332290342/big-data-peeps-at-your-medical-records-to-find-drug-problems> [https://perma.cc/GMT4-ECBD].

4. *Business by Numbers*, ECONOMIST (Sept. 13, 2007), <http://www.economist.com/node/9795140> [https://perma.cc/7YC2-DMYA].

5. Nadya Labi, *Misfortune Teller*, ATLANTIC (Jan.–Feb. 2012), <http://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846> [https://perma.cc/7L72-J5L9].

6. David Lazer et al., *Computational Social Science*, 323 SCI. 721, 722 (2009).

7. Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANN. REV. SOC. 181, 182 (2008).

8. *Id.*

choices, account for a large part of the disparate effects observed.<sup>9</sup> Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society. It can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment.

Algorithms<sup>10</sup> could exhibit these tendencies even if they have not been manually programmed to do so, whether on purpose or by accident. Discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight. Such a possibility has gone unrecognized by most scholars and policy makers, who tend to fear concealed, nefarious intentions or the overlooked effects of human bias or error in hand coding algorithms.<sup>11</sup> Because the discrimination at issue is unintentional, even honest attempts to certify the absence of prejudice on the part of those involved in the data mining process may wrongly confer the imprimatur of impartiality on the resulting decisions. Furthermore, because the mechanism through which data mining may disadvantage protected classes is less obvious in cases of unintentional discrimination, the injustice may be harder to identify and address.

In May 2014, the White House released a report titled *Big Data: Seizing Opportunities, Preserving Values* (Podesta Report), which hinted at the discriminatory potential of big data.<sup>12</sup> The report finds “that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.”<sup>13</sup> It suggests that there may be unintended discriminatory

---

9. See Andrew Grant-Thomas & John A. Powell, *Toward a Structural Racism Framework*, 15 *POVERTY & RACE* 3, 4 (“‘Institutional racism’ was the designation given in the late 1960s to the recognition that, at very least, racism need not be individualist, essentialist or intentional.”).

10. An “algorithm” is a formally specified sequence of logical operations that provides step-by-step instructions for computers to act on data and thus automate decisions. SOLON BAROCAS ET AL., *DATA & CIVIL RIGHTS: TECHNOLOGY PRIMER* (2014), <http://www.datacivilrights.org/pubs/2014-1030/Technology.pdf> [https://perma.cc/X3YX-XHNA]. Algorithms play a role in both automating the discovery of useful patterns in datasets and automating decision making that relies on these discoveries. This Essay uses the term to refer to the latter.

11. See, e.g., Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 *B.C. L. REV.* 93, 101 (2014) (“[H]ousing providers could design an algorithm to predict the [race, gender, or religion] of potential buyers or renters and advertise the properties only to those who [meet certain] profiles.”); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 *WASH. L. REV.* 1, 4 (2014) (“Because human beings program predictive algorithms, their biases and values are embedded into the software’s instructions. . . .”); Danielle Keats Citron, *Technological Due Process*, 85 *WASH. U. L. REV.* 1249, 1254 (2008) (“Programmers routinely change the substance of rules when translating them from human language into computer code.”).

12. EXEC. OFFICE OF THE PRESIDENT, *BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES* (May 2014), [http://www.whitehouse.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_5.1.14\\_final\\_print.pdf](http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf) [https://perma.cc/ZXB4-SDL9].

13. *Id.* (introductory letter).

effects from data mining but does not detail how they might come about.<sup>14</sup> Because the origin of the discriminatory effects remains unexplored, the report's approach does not address the full scope of the problem.

The Podesta Report, as one might expect from the executive branch, seeks to address these effects primarily by finding new ways to enforce existing law. Regarding discrimination, the report primarily recommends that enforcement agencies, such as the Department of Justice, Federal Trade Commission, Consumer Financial Protection Bureau, and Equal Employment Opportunity Commission (EEOC), increase their technical expertise and "develop a plan for investigating and resolving violations of law in such cases."<sup>15</sup>

As this Essay demonstrates, however, existing law largely fails to address the discrimination that can result from data mining. The argument is grounded in Title VII because, of all American antidiscrimination jurisprudence, Title VII has a particularly well-developed set of case law and scholarship. Further, there exists a rapidly emerging field of "work-force science,"<sup>16</sup> for which Title VII will be the primary vehicle for regulation. Under Title VII, it turns out that some, if not most, instances of discriminatory data mining will not generate liability. While the Essay does not show this to be true outside of Title VII itself, the problem is likely not particular to Title VII. Rather, it is a feature of our current approach to antidiscrimination jurisprudence, with its focus on procedural fairness. The analysis will likely apply to other traditional areas of discrimination, such as housing or disability discrimination. Similar tendencies to disadvantage the disadvantaged will likely arise in areas that regulate legitimate economic discrimination, such as credit and insurance.

This Essay proceeds in three Parts. Part I introduces the computer science literature and proceeds through the various steps of solving a problem with data mining: defining the target variable, labeling and collecting the training data, using feature selection, and making decisions on the basis of the resulting model. Each of these steps creates possibilities for a final result that has a disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors. Even in situations where data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes. Finally, Part I notes that data mining poses the additional problem of

---

14. *Id.* at 64 ("This combination of circumstances and technology raises difficult questions about how to ensure that discriminatory effects resulting from automated decision processes, whether intended or not, can be detected, measured, and redressed.")

15. *Id.* at 65.

16. Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), <http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html> [<https://perma.cc/CEL2-P9XB>].

giving data miners the ability to disguise intentional discrimination as accidental.

In Part II, the Essay reviews Title VII jurisprudence as it applies to data mining. Part II discusses both disparate treatment and disparate impact, examining which of the various data mining mechanisms identified in Part I will trigger liability under either Title VII theory. At first blush, either theory is viable. Disparate treatment is viable because data mining systems treat everyone differently; that is their purpose. Disparate impact is also viable because data mining can have various discriminatory effects, even without intent. But as Part II demonstrates, data mining combines some well-known problems in discrimination doctrines with new challenges particular to data mining systems, such that liability for discriminatory data mining will be hard to find. Part II concludes with a discussion of the new problems of proof that arise for intentional discrimination in this context.

Finally, Part III addresses the difficulties reformers would face in addressing the deficiencies found in Part II. These difficulties take two forms: complications internal to the logic of data mining and political and constitutional difficulties external to the problem. Internally, the different steps in a data mining problem require constant subjective and fact-bound judgments, which do not lend themselves to general legislative resolution. Worse, many of these are normative judgments in disguise, about which there is not likely to be consensus. Externally, data mining will force society to explicitly rebalance the two justifications for antidiscrimination law—rooting out intentional discrimination and equalizing the status of historically disadvantaged communities. This is because methods of proof and corrective measures will often require an explicit commitment to substantive remediation rather than merely procedural remedies. In certain cases, data mining will make it simply impossible to rectify discriminatory results without engaging with the question of what level of substantive inequality is proper or acceptable in a given context. Given current political realities and trends in constitutional doctrines, legislation enacting a remedy that results from these discussions faces an uphill battle. To be sure, data mining also has the potential to help reduce discrimination by forcing decisions onto a more reliable empirical foundation and by formalizing decision-making processes, thus limiting the opportunity for individual bias to affect important assessments.<sup>17</sup> In many situations, the introduction of data mining will be a boon to civil rights, even where it fails to root out discrimination altogether, and such efforts should be encouraged. Yet, understanding when and why discrimination persists in cases of data-driven decision making reveals important and sometimes troubling limits to the promise of big data, for which there are no ready solutions.

---

17. Tal Z. Zarsky, *Automated Prediction: Perception, Law, and Policy*, COMM. ACM, Sept. 2012, at 33–35.

## I.

## HOW DATA MINING DISCRIMINATES

Although commentators have ascribed myriad forms of discrimination to data mining,<sup>18</sup> there remains significant confusion over the precise mechanisms that render data mining discriminatory. This Part develops a taxonomy that isolates and explicates the specific technical issues that can give rise to models whose use in decision making may have a disproportionately adverse impact on protected classes. By definition, data mining is *always* a form of statistical (and therefore seemingly rational) discrimination. Indeed, the very point of data mining is to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar. Nevertheless, data mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage. Unlike more subjective forms of decision making, data mining's ill effects are often not traceable to human bias, conscious or unconscious. This Part describes five mechanisms by which these disproportionately adverse outcomes might occur, walking through a sequence of key steps in the overall data mining process.

## A. Defining the "Target Variable" and "Class Labels"

In contrast to those traditional forms of data analysis that simply return records or summary statistics in response to a specific query, data mining attempts to locate statistical relationships in a dataset.<sup>19</sup> In particular, it automates the process of discovering useful patterns, revealing regularities upon which subsequent decision making can rely. The accumulated set of discovered relationships is commonly called a "model," and these models can be employed to automate the process of classifying entities or activities of interest, estimating the value of unobserved variables, or predicting future outcomes.<sup>20</sup> Familiar examples of such applications include spam or fraud detection, credit scoring, and insurance pricing. These examples all involve attempts to determine the status or likely outcome of cases under consideration based solely on access to *correlated* data.<sup>21</sup> Data mining helps identify cases of

---

18. Solon Barocas, *Data Mining and the Discourse on Discrimination*, PROC. DATA ETHICS WORKSHOP (2014), <https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf> [<https://perma.cc/D3LT-GS2X>].

19. See generally Usama Fayyad, *The Digital Physics of Data Mining*, 44 COMM. ACM, Mar. 2001, at 62.

20. More formally, classification deals with discrete outcomes, estimation deals with continuous variables, and prediction deals with both discrete outcomes and continuous variables, but specifically for states or values *in the future*. MICHAEL J. A. BERRY & GORDON S. LINOFF, DATA MINING TECHNIQUES: FOR MARKETING, SALES, AND CUSTOMER RELATIONSHIP MANAGEMENT 8–11 (2004).

21. Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, COMM. ACM, Oct. 2012, at 78–80.

spam and fraud and anticipate default and poor health by treating these states and outcomes as a function of some other set of observed characteristics.<sup>22</sup> In particular, by exposing so-called “machine learning” algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm “learns” which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest.<sup>23</sup>

Two concepts from the machine learning and data mining literature are important here: “target variables” and “class labels.” The outcomes of interest discussed above are known as target variables.<sup>24</sup> While the target variable defines what data miners are looking for, “class labels” divide all possible values of the target variable into mutually exclusive categories.

The proper specification of the target variable is frequently not obvious, and the data miner’s task is to define it. To start, data miners must translate some amorphous problem into a question that can be expressed in more formal terms that computers can parse. In particular, data miners must determine how to solve the problem at hand by translating it into a question about the value of some target variable. The open-endedness that characterizes this part of the process is often described as the “art” of data mining. This initial step requires a data miner to “understand[] the project objectives and requirements from a business perspective [and] then convert[] this knowledge into a data mining problem definition.”<sup>25</sup> Through this necessarily subjective process of translation, data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes.

Problem specification is not a wholly arbitrary process, however. Data mining can only address problems that lend themselves to formalization as questions about the state or value of the target variable. Data mining works exceedingly well for dealing with fraud and spam because these cases rely on extant, binary categories. A given instance either is or is not fraud or spam, and the definitions of fraud or spam are, for the most part, uncontroversial.<sup>26</sup> A computer can then flag or refuse transactions or redirect emails according to

---

22. *Id.*

23. *Id.*

24. COMM. ON THE ANALYSIS OF MASSIVE DATA ET AL., FRONTIERS IN MASSIVE DATA ANALYSIS 101 (2013), [http://www.nap.edu/catalog.php?record\\_id=18374](http://www.nap.edu/catalog.php?record_id=18374) [<https://perma.cc/5DNQ-UFE4>]. The machine learning community refers to classification, estimation, and prediction—the techniques that we discuss in this Essay—as “supervised” learning because analysts must actively specify a target variable of interest. *Id.* at 104. Other techniques known as “unsupervised” learning do not require any such target variables and instead search for general structures in the dataset, rather than patterns specifically related to some state or outcome. *Id.* at 102. Clustering is the most common example of “unsupervised” learning, in that clustering algorithms simply reveal apparent hot spots when plotting the data in some fashion. *Id.* We limit the discussion to supervised learning because we are primarily concerned with the sorting, ranking, and predictions enabled by data mining.

25. PETE CHAPMAN ET AL., CRISP-DM 1.0: STEP-BY-STEP DATA MINING GUIDE 10 (2000).

26. See David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1, 10 (2006).



well-understood distinctions.<sup>27</sup> In these cases, data miners can simply rely on these simple, preexisting categories to define the class labels.

Sometimes, though, defining the target variable involves the creation of *new* classes. Consider credit scoring, for instance. Although now taken for granted, the predicted likelihood of missing a certain number of loan repayments is not a self-evident answer to the question of how to successfully extend credit to consumers.<sup>28</sup> Unlike fraud or spam, “creditworthiness” is an artifact of the problem definition itself. There is no way to directly measure creditworthiness because the very notion of creditworthiness is a function of the particular way the credit industry has constructed the credit issuing and repayment system. That is, an individual’s ability to repay some minimum amount of an outstanding debt on a monthly basis is taken to be a nonarbitrary standard by which to determine in advance and all-at-once whether he is worthy of credit.<sup>29</sup>

Data mining has many uses beyond spam detection, fraud detection, credit scoring, and insurance pricing. As discussed in the introduction, this Essay will focus on the use of data mining in employment decisions. Extending this discussion to employment, then, where employers turn to data mining to develop ways of improving and automating their search for good employees, they face a number of crucial choices.

Like creditworthiness, the definition of a good employee is not a given. “Good” must be defined in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example. When employers mine data for good employees, they are, in fact, looking for employees whose observable characteristics suggest that they would meet or exceed some monthly sales threshold, perform some task in less than a certain amount of time, or remain in their positions for more than a set number of weeks or months. Rather than drawing categorical distinctions along these lines, data mining could also estimate or predict the specific numerical value of sales, production time, or tenure period, enabling employers to rank rather than simply sort employees.

These may seem like eminently reasonable things for employers to want to predict, but they are, by necessity, only part of an array of possible definitions of “good.” An employer may instead attempt to define the target variable in a more holistic way—by, for example, relying on the grades that prior employees have received in annual reviews, which are supposed to reflect

---

27. Though described as a matter of detection, this is really a classification task, where any given transaction or email can belong to one of two possible classes, respectively: fraud or not fraud, or spam or not spam.

28. See generally Martha Ann Poon, *What Lenders See—A History of the Fair Isaac Scorecard*, (2013) (unpublished Ph.D. dissertation, University of California, San Diego), <http://search.proquest.com/docview/1520318884> [https://perma.cc/YD3S-B9N7].

29. Hand, *supra* note 26, at 10.

an overall assessment of performance. These target variable definitions simply inherit the formalizations involved in preexisting assessment mechanisms, which in the case of human-graded performance reviews, may be far less consistent.<sup>30</sup>

Thus, the definition of the target variable and its associated class labels will determine what data mining happens to find. While critics of data mining have tended to focus on inaccurate classifications (false positives and false negatives),<sup>31</sup> as much—if not more—danger resides in the definition of the class label itself and the subsequent labeling of examples from which rules are inferred.<sup>32</sup> While different choices for the target variable and class labels can seem more or less reasonable, valid concerns with discrimination enter at this stage because the different choices may have a greater or lesser adverse impact on protected classes. For example, as later Parts will explain in detail, hiring decisions made on the basis of predicted tenure are much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity. If the turnover rate happens to be systematically higher among members of certain protected classes, hiring decisions based on predicted length of employment will result in fewer job opportunities for members of these groups, even if they would have performed as well as or better than the other applicants the company chooses to hire.

### B. Training Data

As described above, data mining learns by example. Accordingly, what a model learns depends on the examples to which it has been exposed. The data that function as examples are known as “training data”—quite literally, the data that train the model to behave in a certain way. The character of the training data can have meaningful consequences for the lessons that data mining happens to learn. As computer science scholars explain, biased training data leads to discriminatory models.<sup>33</sup> This can mean two rather different things,

---

30. Joseph M. Stauffer & M. Ronald Buckley, *The Existence and Nature of Racial Bias in Supervisory Ratings*, 90 J. APPLIED PSYCHOL. 586, 588–89 (2005) (showing evidence of racial bias in performance evaluations). Nevertheless, devising new target variables can have the salutary effect of forcing decision makers to think much more concretely about the outcomes that justifiably determine whether someone is a “good” employee. The explicit enumeration demanded of data mining thus also presents an opportunity to make decision making more consistent, more accountable, and fairer overall. This, however, requires conscious effort and careful thinking, and is not a natural consequence of adopting data mining.

31. Bruce Schneier, *Data Mining for Terrorists*, SCHNEIER ON SECURITY (Mar. 9, 2006), [https://www.schneier.com/blog/archives/2006/03/data\\_mining\\_for.html](https://www.schneier.com/blog/archives/2006/03/data_mining_for.html) [https://perma.cc/ZW44-N2KR]; Oscar H. Gandy Jr., *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS & INFO. TECH. 29, 39–40 (2010); Mireille Hildebrandt & Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73 MOD. L. REV. 428, 433–35 (2010).

32. See *infra* Part I.B.

33. Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 3, 20 (Bart Custers et al. eds., 2013).

though: (1) if data mining treats cases in which prejudice has played some role as valid examples to learn from, that rule may simply reproduce the prejudice involved in these earlier cases; or (2) if data mining draws inferences from a biased sample of the population, any decision that rests on these inferences may systematically disadvantage those who are under- or overrepresented in the dataset. Both can affect the training data in ways that lead to discrimination, but the mechanisms—improper labeling of examples and biased data collections—are sufficiently distinct that they warrant separate treatment.

### 1. *Labeling Examples*

Labeling examples is the process by which the training data is manually assigned class labels. In cases of fraud or spam, the data miners draw from examples that come pre-labeled: when individual customers report fraudulent charges or mark a message as spam, they are actually labeling transactions and email for the providers of credit and webmail. Likewise, an employer using grades previously given at performance reviews is also using pre-labeled examples.

In certain cases, however, there may not be any labeled data and data miners may have to figure out a way to label examples themselves. This can be a laborious process, and it is frequently fraught with peril.<sup>34</sup> Often the best labels for different classifications will be open to debate. On which side of the creditworthy line does someone who has missed four credit card payments fall, for example?<sup>35</sup> The answer is not obvious. Even where the class labels are uncontested or uncontroversial, they may present a problem because analysts will often face difficult choices in deciding which of the available labels best applies to a particular example. Certain cases may present some, but not all, criteria for inclusion in a particular class.<sup>36</sup> The situation might also work in reverse, where the class labels are insufficiently precise to capture meaningful differences between cases. Such imperfect matches will demand that data miners exercise judgment.

The unavoidably subjective labeling of examples will skew the resulting findings such that any decisions taken on the basis of those findings will characterize all future cases along the same lines. This is true even if such

---

34. Hand, *supra* note 26, at 10–11.

35. *Id.* at 10 (“The classical supervised classification paradigm also takes as fundamental the fact that the classes are well defined. That is, that there is some fixed clear external criterion, which is used to produce the class labels. In many situations, however, this is not the case. In particular, when the classes are defined by thresholding a continuous variable, there is always the possibility that the defining threshold might be changed. Once again, this situation arises in consumer credit, where it is common to define a customer as ‘defaulting’ if they fall three months in arrears with repayments. This definition, however, is not a qualitative one (contrast has a tumor/does not have a tumor) but is very much a quantitative one. It is entirely reasonable that alternative definitions (e.g., four months in arrears) might be more useful if economic conditions were to change.”).

36. *Id.* at 11.

characterizations would seem plainly erroneous to analysts who looked more closely at the individual cases. For all their potential problems, though, the labels applied to the training data must serve as ground truth.<sup>37</sup> Thus, decisions based on discoveries that rest on haphazardly labeled data or data labeled in a systematically, though unintentionally, biased manner will seem valid according to the customary validation methods employed by data miners. So long as prior decisions affected by some form of prejudice serve as examples of *correctly* rendered determinations, data mining will necessarily infer rules that exhibit the same prejudice.

Consider a real-world example from a different context as to how biased data labeling can skew results. St. George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants based on its previous admissions decisions.<sup>38</sup> Those admissions decisions, it turns out, had systematically disfavored racial minorities and women with credentials otherwise equal to other applicants'.<sup>39</sup> In drawing rules from biased prior decisions, St. George's Hospital unknowingly devised an automated process that possessed these very same prejudices. As editors at the *British Medical Journal* noted at the time, "[T]he program was not introducing new bias but merely reflecting that already in the system."<sup>40</sup> Were an employer to undertake a similar plan to automate its hiring decisions by inferring a rule from past decisions swayed by prejudice, the employer would likewise arrive at a decision procedure that simply reproduces the prejudice of prior decision makers. Indeed, automating the process in this way would turn the conscious prejudice or implicit bias of individuals involved in previous decision making into a formalized rule that would systematically alter the prospects of all future applicants. For example, the computer may learn to discriminate against certain female or black applicants if trained on prior hiring decisions in which an employer has consistently rejected jobseekers with degrees from women's or historically black colleges.

Not only can data mining inherit *prior* prejudice through the mislabeling of examples, it can also reflect current prejudice through the ongoing behavior of users taken as inputs to data mining. This is what Professor Latanya Sweeney discovered in a study that found that Google queries for black-sounding names were more likely to return contextual (i.e., key-word triggered)

---

37. *Id.* at 12. Even when evaluating a model, the kinds of subtle mischaracterizations that happen during training will be impossible to detect because most "evaluation data" is just a small subset of the training data that has been withheld during the learning process. Any problems with the training data will be present in the evaluation data.

38. Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 *BRIT. MED. J.* 657, 657 (1988).

39. *Id.* at 657.

40. *Id.*

advertisements for arrest records than those for white-sounding names.<sup>41</sup> Sweeney confirmed that the companies paying for these advertisements had not set out to focus on black-sounding names; rather, the fact that black-sounding names were more likely to trigger such advertisements seemed to be an artifact of the algorithmic process that Google employs to determine which advertisements to display alongside certain queries.<sup>42</sup> Although it is not fully known how Google computes the so-called “quality score” according to which it ranks advertisers’ bids, one important factor is the predicted likelihood, based on historical trends, that users will click on an advertisement.<sup>43</sup> As Sweeney points out, the process “learns over time which [advertisement] text gets the most clicks from viewers [of the advertisement]” and promotes that advertisement in its rankings accordingly.<sup>44</sup> Sweeney posits that this aspect of the process could result in the differential delivery of advertisements that reflect the kinds of prejudice held by those exposed to the advertisements.<sup>45</sup> In attempting to cater to users’ preferences, Google will unintentionally reproduce the existing prejudices that inform users’ choices.

A similar situation could conceivably arise on websites that recommend potential employees to employers, as LinkedIn does through its Talent Match feature.<sup>46</sup> If LinkedIn determines which candidates to recommend based on the demonstrated interest of employers in certain types of candidates, Talent Match will offer recommendations that reflect whatever biases employers happen to exhibit. In particular, if LinkedIn’s algorithm observes that employers disfavor certain candidates who are members of a protected class, Talent Match may decrease the rate at which it recommends these candidates to employers. The recommendation engine would learn to cater to the prejudicial preferences of employers.

There is an old adage in computer science: “garbage in, garbage out.” Because data mining relies on training data as ground truth, when those inputs

---

41. Latanya Sweeney, *Discrimination in Online Ad Delivery*, COMM. ACM, May 2013, at 44, 47 (2013).

42. *Id.* at 48, 52.

43. *Check and Understand Quality Score*, GOOGLE, <https://support.google.com/adwords/answer/2454010?hl=en> [<https://perma.cc/A88T-GF8X>] (last visited July 26, 2014).

44. Sweeney, *supra* note 41, at 52.

45. The fact that black people may be convicted of crimes at a higher rate than nonblack people does not explain why those who search for black-sounding names would be any more likely to click on advertisements that mention an arrest record than those who see the same exact advertisement when they search for white-sounding names. If the advertisement implies, in both cases, that a person of that particular name has an arrest record, as Sweeney shows, the only reason the advertisements keyed to black-sounding names should receive greater attention is if searchers confer greater significance to the fact of prior arrests when the person happens to be black. *Id.* at 53.

46. Dan Woods, *LinkedIn’s Monica Rogati on “What Is a Data Scientist?”*, FORBES (Nov. 27, 2011), <http://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist> [<https://perma.cc/N9HT-BXU3>].

are themselves skewed by bias or inattention, the resulting system will produce results that are at best unreliable and at worst discriminatory.

## 2. Data Collection

Decisions that depend on conclusions drawn from incorrect, partial, or nonrepresentative data may discriminate against protected classes. The individual records that a company maintains about a person might have serious mistakes,<sup>47</sup> the records of the entire protected class of which this person is a member might also have similar mistakes at a higher rate than other groups, and the entire set of records may fail to reflect members of protected classes in accurate proportion to others.<sup>48</sup> In other words, the quality and representativeness of records might vary in ways that correlate with class membership (e.g., institutions might maintain systematically less accurate, precise, timely, and complete records for certain classes of people). Even a dataset with individual records of consistently high quality can suffer from statistical biases that fail to represent different groups in accurate proportions. Much attention has focused on the harms that might befall individuals whose records in various commercial databases are error ridden.<sup>49</sup> Far less consideration, however, has been paid to the systematic disadvantage that members of protected classes may suffer from being miscounted and, as a result, misrepresented in the evidence base.

Recent scholarship has begun to stress this point. Jonas Lerman, for example, worries about “the nonrandom, systemic omission of people who live on big data’s margins, whether due to poverty, geography, or lifestyle, and whose lives are less ‘datafied’ than the general population’s.”<sup>50</sup> Professor Kate Crawford has likewise warned that “[b]ecause not all data is created or even collected equally, there are ‘signal problems’ in big-data sets—dark zones or shadows where some citizens and communities are overlooked or

---

47. Data quality is a topic of lively practical and philosophical debate. *See, e.g.*, Luciano Floridi, *Information Quality*, 26 PHIL. & TECH. 1 (2013); Richard Y. Wang & Diane M. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers*, 12 J. MGMT. INFO. SYS. 5 (1996). The components of data quality have been thought to include accuracy, precision, completeness, consistency, validity, and timeliness, though this catalog of features is far from settled. *See generally* LARRY P. ENGLISH, INFORMATION QUALITY APPLIED (2009).

48. *Cf.* Zeynep Tufekci, *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*, EIGHTH INT’L AAAI CONF. WEBLOGS & SOC. MEDIA (2014), <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8062/8151> [<https://perma.cc/G4G7-2VZ8>].

49. *See, e.g.*, FED. TRADE COMM’N, REPORT TO CONGRESS UNDER SECTION 319 OF THE FAIR AND ACCURATE CREDIT TRANSACTIONS ACT OF 2003 A-4 (2012) (finding that nearly 20 percent of consumers had an error in one or more of their three credit reports and that 5.4 percent of consumers had errors that could result in less favorable loan terms).

50. Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 57 (2013).

underrepresented.”<sup>51</sup> Errors of this sort may befall historically disadvantaged groups at higher rates because they are less involved in the formal economy and its data-generating activities, have unequal access to and relatively less fluency in the technology necessary to engage online, or are less profitable customers or important constituents and therefore less interesting as targets of observation.<sup>52</sup> Not only will the quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well represented in datasets, skewing conclusions that may be drawn from an analysis of the data.

As an illustrative example, Crawford points to Street Bump, an application for Boston residents that takes advantage of accelerometers built into smart phones to detect when drivers ride over potholes.<sup>53</sup> While Crawford praises the cleverness and cost-effectiveness of this passive approach to reporting road problems, she rightly warns that whatever information the city receives from Street Bump will be biased by the uneven distribution of smartphones across populations in different parts of the city.<sup>54</sup> In particular, systematic differences in smartphone ownership will very likely result in the underreporting of road problems in the poorer communities where protected groups disproportionately congregate.<sup>55</sup> If the city were to rely on this data to determine where it should direct its resources, it would only further underserve these communities. Indeed, the city would discriminate against those who lack the capability to report problems as effectively as wealthier residents with smartphones.<sup>56</sup>

A similar dynamic could easily apply in an employment context if members of protected classes are unable to report their interest in and qualification for jobs listed online as easily or effectively as others due to systematic differences in Internet access. The EEOC has established a program called “Eradicating Racism & Colorism from Employment” (E-RACE) that aims, at least in part, to prevent this sort of discrimination from occurring due

---

51. Kate Crawford, *Think Again: Big Data*, FOREIGN POL’Y (May 10, 2013), [http://www.foreignpolicy.com/articles/2013/05/09/think\\_again\\_big\\_data](http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data) [https://perma.cc/S9ZA-XEXH].

52. *See id.*; Lerman, *supra* note 50, at 57.

53. Crawford, *supra* note 51 (explaining that a sudden movement suggesting a broken road will automatically prompt the phone to report the location to the city).

54. *Id.*

55. *See id.*

56. This is, of course, a more general problem with representative democracy. For a host of reasons, the views and interests of the poor are relatively less well represented in the political process. *See, e.g.*, Larry M. Bartels, *Economic Inequality and Political Representation*, in THE UNSUSTAINABLE AMERICAN STATE 167 (Lawrence Jacobs & Desmond King eds., 2009); MARTIN GILENS, AFFLUENCE AND INFLUENCE: ECONOMIC INEQUALITY AND POLITICAL POWER IN AMERICA (2012). The worry here, as expressed by Crawford, is that, for all its apparent promise, data mining may further obfuscate or legitimize these dynamics rather than overcome them.

to an employer's desire for high-tech hiring, such as video résumés.<sup>57</sup> E-RACE not only attempts to lower the barriers that would disproportionately burden applicants who belong to a protected class, but also ensures that employers do not develop an inaccurate impression of the incidence of qualified and interested candidates from these communities. If employers were to rely on tallies of high-tech candidates to direct their recruiting efforts, for example, any count affected by a reporting bias could have adverse consequences for specific populations systematically underrepresented in the dataset. Employers would deny equal attention to those who reside in areas incorrectly pegged as having a relatively lower concentration of qualified candidates.

Additional and even more severe risks may reside in the systematic omission of members of protected classes from such datasets. The Street Bump and Internet job application examples only discuss decisions that depend on raw tallies, rather than datasets from which decision makers want to draw generalizations and generate predictions. But data mining is especially sensitive to statistical bias because data mining helps to discover patterns that organizations tend to treat as generalizable findings even though the analyzed data only includes a partial sample from a circumscribed period. To ensure that data mining reveals patterns that hold true for more than the particular sample under analysis, the sample must be proportionally representative of the entire population, even though the sample, by definition, does not include every case.<sup>58</sup>

If a sample includes a disproportionate representation of a particular class (more or less than its actual incidence in the overall population), the results of an analysis of that sample may skew in favor of or against the over- or underrepresented class. While the representativeness of the data is often simply assumed, this assumption is rarely justified and is “perhaps more often incorrect than correct.”<sup>59</sup> Data gathered for routine business purposes tend to lack the rigor of social scientific data collection.<sup>60</sup> As Lerman points out, “Businesses may ignore or undervalue the preferences and behaviors of

---

57. *Why Do We Need E-RACE?*, EQUAL EMPLOY. OPPORTUNITY COMM'N, [http://www1.eeoc.gov/eeoc/initiatives/e-race/why\\_e-race.cfm](http://www1.eeoc.gov/eeoc/initiatives/e-race/why_e-race.cfm) [<https://perma.cc/S3GY-2MD6>] (last visited Mar. 1, 2013). Due to the so-called “digital divide,” communities underserved by residential Internet access rely heavily on mobile phones for connectivity and thus often have trouble even uploading and updating traditional résumés. Kathryn Zickuhr & Aaron Smith, *Digital Differences*, PEW RES. CTR. (Apr. 13, 2012), <http://www.pewinternet.org/2012/04/13/digital-differences> [<https://perma.cc/S545-42GY>] (“Among smartphone owners, young adults, minorities, those with no college experience, and those with lower household income levels are more likely than other groups to say that their phone is their main source of internet access.”).

58. Data mining scholars have devised ways to address this known problem, but applying these techniques is far from trivial. See Sinno Jialin Pan & Qiang Yang, *A Survey on Transfer Learning*, 22 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENG'G 1345, 1354–56 (2010).

59. Hand, *supra* note 26, at 7.

60. David Lazer, *Big Data and Cloning Headless Frogs*, COMPLEXITY & SOC. NETWORKS BLOG (Feb. 16, 2014), [https://web.archive.org/web/20140711164511/http://blogs.iq.harvard.edu/netgov/2014/02/big\\_data\\_and\\_cloning\\_headless.html](https://web.archive.org/web/20140711164511/http://blogs.iq.harvard.edu/netgov/2014/02/big_data_and_cloning_headless.html) [<https://perma.cc/TQ9A-TP2Z>].



consumers who do not shop in ways that big data tools can easily capture, aggregate, and analyze.”<sup>61</sup>

In the employment context, even where a company performs an analysis of the data from its entire population of employees—avoiding the apparent problem of even having to select a sample—the organization must assume that its future applicant pool will have the same degree of variance as its current employee base. An organization’s tendency, however, to perform such analyses in order to *change* the composition of their employee base should put the validity of this assumption into immediate doubt. The potential effect of this assumption is the future mistreatment of individuals predicted to behave in accordance with the skewed findings derived from the biased sample. Worse, these results may lead to decision procedures that limit the future contact an organization will have with specific groups, skewing still further the sample upon which subsequent analyses will be performed.<sup>62</sup> Limiting contact with specific populations on the basis of unsound generalizations may deny members of these populations the opportunity to prove that they buck the apparent trend.

*Overrepresentation* in a dataset can also lead to disproportionately high adverse outcomes for members of protected classes. Consider an example from the workplace: managers may devote disproportionate attention to monitoring the activities of employees who belong to a protected class and consequently observe mistakes and transgressions at systematically higher rates than others, in part because these managers fail to subject others who behave similarly to the same degree of scrutiny. Not only does this provide managers with justification for their prejudicial suspicions, but it also generates evidence that overstates the relative incidence of offenses by members of these groups. Where subsequent managers who hold no such prejudicial suspicions cannot observe everyone equally, they may rely on this evidence to make predictions about where to focus their attention in the future and thus further increase the disproportionate scrutiny that they place on protected classes.

The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making.

---

61. Lerman, *supra* note 50, at 59.

62. Practitioners, particularly those involved in credit scoring, are well aware that they do not know how the person purposefully passed over would have behaved if he had been given the opportunity. Practitioners have developed methods to correct for this bias (which, in the case of credit scoring, they refer to as reject inference). See, e.g., Jonathan Crook & John Banasik, *Does Reject Inference Really Improve the Performance of Application Scoring Models?*, 28 J. BANKING & FIN. 857 (2004).

### C. Feature Selection

Through a process called “feature selection,” organizations—and the data miners that work for them—make choices about what attributes they observe and subsequently fold into their analyses.<sup>63</sup> These decisions can also have serious implications for the treatment of protected classes if those factors that better account for pertinent statistical variation among members of a protected class are not well represented in the set of selected features.<sup>64</sup> Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve.

This problem arises because data are necessarily reductive representations of an infinitely more specific real-world object or phenomenon.<sup>65</sup> These representations may fail to capture enough detail to allow for the discovery of crucial points of contrast. Increasing the resolution and range of the analysis may still fail to capture the mechanisms that account for different outcomes because such mechanisms may not lend themselves to exhaustive or effective representation in the data, if such representations even exist. As Professors Toon Calders and Indrė Žliobaitė explain, “[I]t is often impossible to collect all the attributes of a subject or take all the environmental factors into account with a model.”<sup>66</sup> While these limitations lend credence to the argument that a dataset can never fully encompass the full complexity of the individuals it seeks to represent, they do not reveal the inherent inadequacy of representation as such.

At issue, really, are the coarseness and comprehensiveness of the criteria that permit statistical discrimination and the uneven rates at which different groups happen to be subject to erroneous determinations. Crucially, these erroneous and potentially adverse outcomes are artifacts of statistical reasoning rather than prejudice on the part of decision makers or bias in the composition of the dataset. As Professor Frederick Schauer explains, decision makers that rely on statistically sound but nonuniversal generalizations “are being simultaneously rational and unfair” because certain individuals are “actuarially saddled” by statistically sound inferences that are nevertheless inaccurate.<sup>67</sup>

---

63. FEATURE EXTRACTION, CONSTRUCTION AND SELECTION 71–72 (Huan Liu & Hiroshi Motoda eds., 1998).

64. Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 33, at 43, 46 (“[T]he selection of attributes by which people are described in [a] database may be incomplete.”).

65. Annamarie Carusi, *Data as Representation: Beyond Anonymity in E-Research Ethics*, 1 INT’L J. INTERNET RES. ETHICS 37, 48–61 (2008).

66. Calders & Žliobaitė, *supra* note 64, at 47.

67. FREDERICK SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 3–7 (2006). Insurance offers the most obvious example of this: the rate that a person pays for car insurance, for

Obtaining information that is sufficiently rich to permit precise distinctions can be expensive. Even marginal improvements in accuracy may come at significant practical costs and may justify a less granular and encompassing analysis.<sup>68</sup>

To take an obvious example from the employment context, hiring decisions that consider academic credentials tend to assign enormous weight to the reputation of the college or university from which an applicant has graduated, even though such reputations may communicate very little about the applicant's job-related skills and competencies.<sup>69</sup> If equally competent members of protected classes happen to graduate from these colleges or universities at disproportionately low rates, decisions that turn on the credentials conferred by these schools, rather than some set of more specific qualities that more accurately sort individuals, will incorrectly and systematically discount these individuals. Even if employers have a rational incentive to look beyond credentials and focus on criteria that allow for more precise and more accurate determinations, they may continue to favor credentials because they communicate pertinent information at no cost to the employer.<sup>70</sup>

Similar dynamics seem to account for the practice known as "redlining,"<sup>71</sup> in which financial institutions employ especially general criteria to draw distinctions between subpopulations (i.e., the neighborhood in which individuals happen to reside), despite the fact that such distinctions fail to capture significant variation within each subpopulation that would result in a different assessment for certain members of these groups. While redlining in America is well known to have had its basis in racial animus and prejudice,<sup>72</sup> decision makers operating in this manner may attempt to justify their behavior by pointing to the cost efficiency of relying on easily accessible information. In other words, decision makers can argue that they are willing to tolerate higher rates of erroneous determinations for certain groups because the benefits

---

instance, is determined by the way other people with similar characteristics happen to drive, even if the person is a better driver than those who resemble him on the statistically pertinent dimensions.

68. Kasper Lippert-Rasmussen, "We Are All Different": *Statistical Discrimination and the Right to Be Treated as an Individual*, 15 J. ETHICS 47, 54 (2011) ("[O]btaining information is costly, so it is morally justified, all things considered, to treat people on the basis of statistical generalizations even though one knows that, in effect, this will mean that one will treat some people in ways, for better or worse, that they do not deserve to be treated."); see also Brian Dalessandro, Claudia Perlich & Troy Raeder, *Bigger Is Better, but at What Cost?: Estimating the Economic Value of Incremental Data Assets*, 2 BIG DATA 87 (2014).

69. See Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. TIMES (Apr. 28, 2013), <http://www.nytimes.com/2013/04/28/technology/how-big-data-is-playing-recruiter-for-specialized-workers.html> [<https://perma.cc/DC7A-W2B5>].

70. As one commentator has put it in contemplating data-driven hiring, "Big Data has its own bias. . . . You measure what you can measure." *Id.*

71. See generally DAVID M. P. FREUND, *COLORED PROPERTY: STATE POLICY AND WHITE RACIAL POLITICS IN SUBURBAN AMERICA* (2010).

72. *Id.*

derived from more granular data—and thus better accuracy—do not justify the costs. Of course, it may be no coincidence that such cost-benefit analyses seem to justify treating groups composed disproportionately of members of protected classes to systematically less accurate determinations.<sup>73</sup> Redlining is illegal because it can systematically discount entire areas composed primarily of members of a protected class, despite the presence of some qualified candidates.<sup>74</sup>

Cases of so-called rational racism are really just a special instance of this more general phenomenon—one in which race happens to be taken into consideration explicitly. In such cases, decision makers take membership in a protected class into account, even if they hold no prejudicial views, because such membership seems to communicate relevant information that would be difficult or impossible to obtain otherwise. Accordingly, the persistence of distasteful forms of discrimination may be the result of a lack of information, rather than a continued taste for discrimination.<sup>75</sup> Professor Lior Strahilevitz has argued, for instance, that when employers lack access to criminal records, they may consider race in assessing an applicant's likelihood of having a criminal record because there are statistical differences in the rates at which members of different racial groups have been convicted of crimes.<sup>76</sup> In other words, employers fall back on more immediately available and coarse features when they cannot access more specific or verified information.<sup>77</sup> Of course, as Strahilevitz points out, race is a highly imperfect basis upon which to predict an individual's criminal record, despite whatever differences may exist in the rates at which members of different racial groups have been convicted of crimes, because it is too coarse as an indicator.<sup>78</sup>

---

73. While animus was likely the main motivating factor for redlining, the stated rationales were economic and about housing value. See DOUGLAS S. MASSEY & NANCY A. DENTON, *AMERICAN APARTHEID: SEGREGATION AND THE MAKING OF THE UNDERCLASS* 51–52 (1993). Redlining persists today and may actually be motivated by profit, but it has the same deleterious effects. See Rachel L. Swarns, *Biased Lending Evolves, and Blacks Face Trouble Getting Mortgages*, N.Y. TIMES (Oct. 30 2015), <http://www.nytimes.com/2015/10/31/nyregion/udson-city-bank-settlement.html> [https://perma.cc/P4YX-NTT9].

74. See *Nationwide Mut. Ins. Co. v. Cisneros*, 52 F.3d 1351, 1359 (6th Cir. 1995) (holding that the Fair Housing Act prohibited redlining in order “to eliminate the discriminatory business practices which might prevent a person economically able to do so from purchasing a house regardless of his race”); *NAACP v. Am. Family Mut. Ins. Co.*, 978 F.2d 287, 300 (7th Cir. 1992).

75. See generally Andrea Romei & Salvatore Ruggieri, *Discrimination Data Analysis: A Multi-Disciplinary Bibliography*, in *DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY*, *supra* note 33, at 109, 120.

76. Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363, 364 (2008).

77. *Id.* This argument assumes that criminal records are relevant to employment, which is often not true. See *infra* text accompanying note 175.

78. Strahilevitz, *supra* note 76, at 364; see also *infra* Part II.A. The law holds that decision makers should refrain from considering membership in a protected class even if statistical evidence seems to support certain inferences on that basis. The prohibition does not depend on whether decision

#### D. Proxies

Cases of decision making that do not artificially introduce discriminatory effects into the data mining process may nevertheless result in systematically less favorable determinations for members of protected classes. This is possible when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job—as formalized in some fashion—may also sort individuals according to class membership.

In certain cases, there may be an obvious reason for this. Just as “mining from historical data may . . . discover traditional prejudices that are endemic in reality (i.e., taste-based discrimination),” so, too, may data mining “discover patterns of lower performances, skills or capacities of protected-by-law groups.”<sup>79</sup> These discoveries not only reveal the simple fact of inequality, but they also reveal that these are inequalities in which members of protected classes are frequently in the relatively less favorable position. This has rather obvious implications: if features held at a lower rate by members of protected groups nevertheless possess relevance in rendering legitimate decisions, such decisions will necessarily result in systematically less favorable determinations for these individuals. For example, by conferring greater attention and opportunities to employees that they predict will prove most competent at some task, employers may find that they subject members of protected groups to consistently disadvantageous treatment because the criteria that determine the attractiveness of employees happen to be held at systematically lower rates by members of these groups.<sup>80</sup>

Decision makers do not necessarily intend this disparate impact because they hold prejudicial beliefs; rather, their reasonable priorities as profit seekers unintentionally recapitulate the inequality that happens to exist in society. Furthermore, this may occur even if proscribed criteria have been removed from the dataset, the data are free from latent prejudice or bias, the features are especially granular and diverse, and the only goal is to maximize classificatory or predictive accuracy. The problem stems from what researchers call “redundant encodings,” cases in which membership in a protected class happens to be encoded in other data.<sup>81</sup> This occurs when a particular piece of data or certain values for that piece of data are highly correlated with

---

makers can gain (easy or cheap) access to alternative criteria that hold greater predictive value. *See* Grutter v. Bollinger, 539 U.S. 306, 326 (2003).

79. Romei & Ruggieri, *supra* note 75, at 121.

80. Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Techniques for Discrimination-Free Predictive Models*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 33, at 223–24.

81. Cynthia Dwork et al., *Fairness Through Awareness*, 3 PROC. INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214 app. at 226 (2012) (“Catalog of Evils”).

membership in specific protected classes. Data's significant statistical relevance to the decision at hand helps explain why data mining can result in seemingly discriminatory models even when its only objective is to ensure the greatest possible accuracy for its determinations. If there is a disparate distribution of an attribute, a more precise form of data mining will be more likely to capture that distribution. Better data and more features will simply come closer to exposing the exact extent of inequality.

### *E. Masking*

Data mining could also breathe new life into traditional forms of intentional discrimination because decision makers with prejudicial views can mask their intentions by exploiting each of the mechanisms enumerated above. Stated simply, any form of discrimination that happens unintentionally can also be orchestrated intentionally. For instance, decision makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes.<sup>82</sup> They could likewise attempt to preserve the known effects of prejudice in prior decision making by insisting that such decisions constitute a reliable and impartial set of examples from which to induce a decision-making rule. And decision makers could intentionally rely on features that only permit coarse-grained distinction making—distinctions that result in avoidably higher rates of erroneous determinations for members of a protected class. In denying themselves finer-grained detail, decision makers would be able to justify writing off entire groups composed disproportionately of members of protected classes. A form of digital redlining, this decision masks efforts to engage in intentional discrimination by abstracting to a level of analysis that fails to capture lower level variations. As a result, certain members of protected classes might not be seen as attractive candidates. Here, prejudice rather than some legitimate business reason (such as cost) motivates decision makers to intentionally restrict the particularity of their decision making to a level that can only paint in avoidably broad strokes. This condemns entire groups, composed disproportionately of members of protected classes, to systematically less favorable treatment.

Because data mining holds the potential to infer otherwise unseen attributes, including those traditionally deemed sensitive,<sup>83</sup> it can indirectly determine individuals' membership in protected classes and unduly discount, penalize, or exclude such people accordingly. In other words, data mining could grant decision makers the ability to distinguish and disadvantage members of protected classes even if those decision makers do not have access to explicit information about individuals' class membership. Data mining could

---

82. See *id.* (discussing the “[s]elf-fulfilling prophecy”).

83. See Solon Barocas, *Leaps and Bounds: Toward a Normative Theory of Inferential Privacy* 9 (Nov. 11, 2015) (in-progress and unpublished manuscript) (on file with authors).

instead help to pinpoint reliable proxies for such membership and thus place institutions in the position to automatically sort individuals into their respective class without ever having to learn these facts directly.<sup>84</sup> The most immediate implication is that institutions could employ data mining to circumvent the barriers, both practical and legal, that have helped to withhold individuals' protected class membership from consideration.

Additionally, data mining could provide cover for intentional discrimination of this sort because the process conceals the fact that decision makers determined and considered the individual's class membership. The worry, then, is not simply that data mining introduces novel ways for decision makers to satisfy their taste for illegal discrimination; rather, the worry is that it may mask actual cases of such discrimination.<sup>85</sup> Although scholars, policy makers, and lawyers have long been aware of the dangers of masking,<sup>86</sup> data mining significantly enhances the ability to conceal acts of intentional discrimination by finding ever more remote and complex proxies for proscribed criteria.<sup>87</sup>

Intentional discrimination and its masking have so far garnered disproportionate attention in discussions of data mining,<sup>88</sup> often to the exclusion of issues arising from the many forms of unintentional discrimination described above. While data mining certainly introduces novel ways to discriminate intentionally and to conceal those intentions, most cases of employment discrimination are already sufficiently difficult to prove; employers motivated by conscious prejudice would have little to gain by pursuing these complex and costly mechanisms to further mask their intentions.<sup>89</sup> When it comes to data mining, unintentional discrimination is the more pressing concern because it is likely to be far more common and easier to overlook.

---

84. *Id.* at 9–13.

85. Data miners who wish to discriminate can do so using relevant or irrelevant criteria. Either way the intent would make the action “masking.” If an employer masked using highly relevant data, litigation arising from it likely would be tried under a “mixed-motive” framework, which asks whether the same action would have been taken without the intent to discriminate. *See infra* Part II.A.

86. *See, e.g.*, Custers, *supra* note 33, at 9–10.

87. *See* Barocas, *supra* note 83.

88. *See, e.g.*, Alistair Croll, *Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It*, SOLVE FOR INTERESTING (July 31, 2012, 12:40 PM), <http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it> [<https://perma.cc/BS8S-6T7S>]. This post generated significant online chatter immediately upon publication and has become one of the canonical texts in the current debate. It has also prompted a number of responses from scholars. *See, e.g.*, Anders Sandberg, *Asking the Right Questions: Big Data and Civil Rights*, PRAC. ETHICS (Aug. 16, 2012), <http://blog.practicaethics.ox.ac.uk/2012/08/asking-the-right-questions-big-data-and-civil-rights> [<https://perma.cc/NC36-NBZN>].

89. *See* Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1177 (1995).

## II.

## TITLE VII LIABILITY FOR DISCRIMINATORY DATA MINING

Current antidiscrimination law is not well equipped to address the cases of discrimination stemming from the problems described in Part I. This Part considers how Title VII might apply to these cases. Other antidiscrimination laws, such as the Americans with Disabilities Act, will exhibit differences in specific operation, but the main thrust of antidiscrimination law is fairly consistent across regimes, and Title VII serves as an illustrative example.<sup>90</sup>

An employer sued under Title VII may be found liable for employment discrimination under one of two theories of liability: disparate treatment and disparate impact.<sup>91</sup> Disparate treatment comprises two different strains of discrimination: (1) formal disparate treatment of similarly situated people and (2) intent to discriminate.<sup>92</sup> Disparate impact refers to policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes.<sup>93</sup> Disparate impact is not concerned with the intent or motive for a policy; where it applies, the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result.<sup>94</sup>

Liability under Title VII for discriminatory data mining will depend on the particular mechanism by which the inequitable outcomes are generated. This Part explores the disparate treatment and disparate impact doctrines and analyzes which mechanisms could generate liability under each theory.

A. *Disparate Treatment*

Disparate treatment recognizes liability for both explicit formal classification and intentional discrimination.<sup>95</sup> Formal discrimination, in which membership in a protected class is used as an input to the model, corresponds to an employer classifying employees or potential hires according to membership in a protected class and differentiating them on that basis. Formal

---

90. The biggest difference between the Americans with Disabilities Act and Title VII is the requirement that an employer make “reasonable accommodations” for disabilities. 42 U.S.C. § 12112(b)(5) (2012). But some scholars have argued that even this difference is illusory and that accommodations law is functionally similar to Title VII, though worded differently. See Samuel R. Bagenstos, “Rational Discrimination,” *Accommodation, and the Politics of (Disability) Civil Rights*, 89 VA. L. REV. 825, 833 & n.15 (2003) (comparing accommodations law to disparate treatment); Christine Jolls, *Antidiscrimination and Accommodation*, 115 HARV. L. REV. 642, 652 (2001) (comparing accommodations law to disparate impact).

91. See 42 U.S.C. § 2000e; *Ricci v. DeStefano*, 557 U.S. 557, 577 (2009).

92. Richard A. Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1351 n.56 (2010) (explaining that, for historical reasons, disparate treatment became essentially “not-disparate-impact” and now we rarely notice the two different embedded theories).

93. See *Griggs v. Duke Power Co.*, 401 U.S. 424, 430 (1971).

94. 42 U.S.C. § 2000e-2(k).

95. *Id.* § 2000e-2(a), (k); see Primus, *supra* note 92, at 1350–51 n.56.



discrimination covers both the straightforward denial of opportunities based on protected class membership and the use of rational racism.<sup>96</sup> In traditional contexts, rational racism is considered rational because there are cases in which its users believe it is an accurate, if coarse-grained, proxy—or at least the best available one in a given situation.<sup>97</sup> In the world of data mining, though, that need not be the case. Even if membership in a protected class were specified as an input, the eventual model that emerges could see it as the least significant feature. In that case, there would be no discriminatory effect, but there would be a disparate treatment violation, because considering membership in a protected class as a potential proxy is a legal classificatory harm in itself.<sup>98</sup>

Formal liability does not correspond to any particular discrimination mechanism within data mining; it can occur equally well in any of them. Because classification itself can be a legal harm, irrespective of the effect,<sup>99</sup> the same should be true of using protected class as an input to a system for which the entire purpose is to build a classificatory model.<sup>100</sup> The irony is that the use of protected class as an input is usually irrelevant to the outcome in terms of discriminatory effect, at least given a large enough number of input features. The target variable will, in reality, be correlated to the membership in a protected class somewhere between 0 percent and 100 percent. If the trait is perfectly uncorrelated, including membership in the protected class as an input will not change the output, and there will be no discriminatory effect.<sup>101</sup> On the other end of the spectrum, where membership in the protected class is perfectly predictive of the target variable, the fact will be redundantly encoded in the other data. The only way using membership in the protected class as an explicit feature will change the outcome is if the information is otherwise not rich enough to detect such membership. Membership in the protected class will prove relevant to the exact extent it is already redundantly encoded. Given a rich enough set of features, the chance that such membership is redundantly encoded approaches certainty. Thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input. Formal discrimination therefore should have no bearing whatsoever on the

---

96. Michelle R. Gomez, *The Next Generation of Disparate Treatment: A Merger of Law and Social Science*, 32 REV. LITIG. 553, 562 (2013).

97. Strahilevitz, *supra* note 76, at 365–67.

98. Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 504 (2003).

99. See Jed Rubenfeld, *Affirmative Action*, 107 YALE L.J. 427, 433 (1997) (discussing “[c]lassificationism”); Primus, *supra* note 98, at 504, 567–68 (discussing expressive harms).

100. Membership in a protected class is still a permissible input to a holistic determination when the focus is diversity, but where classification is the goal, such as here, it is not. See *Grutter v. Bollinger*, 539 U.S. 306, 325 (2003) (noting that “diversity is a compelling state interest” that can survive strict scrutiny).

101. That is, not counting any expressive harm that might come from classification by protected class.

outcome of the model. Additionally, by analyzing the data, an employer could probabilistically determine an employee's membership in that same protected class, if the employer did indeed want to know.

To analyze intentional discrimination other than mere formal discrimination, a brief description of disparate treatment doctrine is necessary. A Title VII disparate treatment case will generally proceed under either the *McDonnell-Douglas* burden-shifting scheme or the *Price-Waterhouse* "mixed motive" regime.<sup>102</sup> Under the *McDonnell-Douglas* framework, the plaintiff who has suffered an adverse employment action has the initial responsibility to establish a prima facie case of discrimination by demonstrating that a similarly situated person who is not a member of a protected class would not have suffered the same fate.<sup>103</sup> This can be shown with circumstantial evidence of discriminatory intent, such as disparaging remarks made by the employer or procedural irregularities in promotion or hiring; only very rarely will an employer openly admit to discriminatory conduct. If the plaintiff successfully demonstrates that the adverse action treated protected class members differently, then the burden shifts to the defendant-employer to offer a legitimate, nondiscriminatory basis for the decision. The defendant need not prove the reason is true; his is only a burden of production.<sup>104</sup> Once the defendant has offered a nondiscriminatory alternative, the ultimate burden of persuasion falls to the plaintiff to demonstrate that the proffered reason is pretextual.<sup>105</sup>

In the data mining context, liability for masking is clear as a theoretical matter, no matter which mechanism for discrimination is employed. The fact that it is accomplished algorithmically does not make it less of a disparate treatment violation, as the entire idea of masking is pretextual. In fact, in the traditional, non-data mining context, the word masking has occasionally been used to refer to pretext.<sup>106</sup> Like in any disparate treatment case, however, proof will be difficult to come by, something even truer for masking.<sup>107</sup>

---

102. *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973); *Price Waterhouse v. Hopkins*, 490 U.S. 228 (1989).

103. This is similar to the computer science definition of discrimination. Calders & Žliobaitė, *supra* note 64, at 49. ("A classifier discriminates with respect to a sensitive attribute, e.g. gender, if for two persons which only differ by their gender (and maybe some characteristics irrelevant for the classification problem at hand) that classifier predicts different labels.").

104. *St. Mary's Honor Ctr. v. Hicks*, 509 U.S. 502, 507 (1993).

105. *Id.*

106. See *Keyes v. Sec'y of the Navy*, 853 F.2d 1016, 1026 (1st Cir. 1988) (explaining that it is the plaintiff's burden to show that the proffered reasons for hiring an alternative were "pretexts aimed at masking sex or race discrimination"); Custers, *supra* note 33, at 9–10; Megan Whitehill, *Better Safe than Subjective: The Problematic Intersection of Prehire Social Networking Checks and Title VII Employment Discrimination*, 85 TEMP. L. REV. 229, 250 (2012) (referring to "[m]asking [p]retext" in the third stage of *McDonnell-Douglas* framework).

107. See *supra* Part I.E. This is a familiar problem to antidiscrimination law, and it is often cited as one of the rationales for disparate impact liability in the first place—to "smoke out" intentional invidious discrimination. See *infra* Part III.B.

The *McDonnell-Douglas* framework operates on a presumption that if the rationale that the employer has given is found to be untrue, the employer must be hiding his “true” discriminatory motive.<sup>108</sup> Because the focus of the *McDonnell-Douglas* framework is on pretext and cover-up, it can only address conscious, willful discrimination.<sup>109</sup> Under the *McDonnell-Douglas* framework, a court must find either that the employer *intended* to discriminate or did not discriminate at all.<sup>110</sup> Thus, unintentional discrimination will not lead to liability.

A Title VII disparate treatment case can also be tried under the mixed-motive framework, first recognized in *Price Waterhouse v. Hopkins*<sup>111</sup> and most recently modified by *Desert Palace, Inc. v. Costa*.<sup>112</sup> In the mixed-motive framework, a plaintiff need not demonstrate that the employer’s nondiscriminatory rationale was pretextual, but merely that discrimination was a “motivating factor” in the adverse employment action.<sup>113</sup> As a practical matter, this means that the plaintiff must show that the same action would not have been taken absent the discriminatory motive.<sup>114</sup> As several commentators

---

108. *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 805 (1973) (The plaintiff “must be given a full and fair opportunity to demonstrate by competent evidence that the presumptively valid reasons for his rejection were in fact a coverup for a racially discriminatory decision”). While, as a theoretical matter, the plaintiff must prove that the employer’s reason was a pretext for discrimination specifically, the Supreme Court has held that a jury can reasonably find that the fact that an employer had only a pretextual reason to fall back on is itself circumstantial evidence of discrimination. *Hicks*, 509 U.S. at 511 (“The factfinder’s disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the prima facie case, suffice to show intentional discrimination.”).

109. See Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. C.R.-C.L. L. REV. 91, 114 (2003) (“Presuming that individuals know the real reason for their actions, the pretext model of disparate treatment provides that an employer can be held to have discriminated when the plaintiff establishes a minimal prima facie case and shows that the reason given for the adverse decision is unworthy of credence.”); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 458 (2001); see also Melissa Hart, *Subjective Decisionmaking and Unconscious Discrimination*, 56 ALA. L. REV. 741, 749–50 (2005) (critiquing the courts’ requirement of proving employer “dishonesty,” but suggesting that, absent this requirement, Title VII could handle unconscious discrimination without altering the law).

110. Krieger, *supra* note 89, at 1170.

111. 490 U.S. 228 (1989).

112. 539 U.S. 90 (2003).

113. 42 U.S.C. § 2000e-2(m) (2012); *Desert Palace*, 539 U.S. at 101 (“In order to obtain [a mixed-motive jury instruction], a plaintiff need only present sufficient evidence for a reasonable jury to conclude, by a preponderance of the evidence, that ‘race, color, religion, sex, or national origin was a motivating factor for any employment practice.’”). The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making.

114. Charles A. Sullivan, *Disparate Impact: Looking Past the Desert Palace Mirage*, 47 WM. & MARY L. REV. 911, 914–16, 916 n.20 (2005); see also Krieger, *supra* note 89, at 1170–72; D. Don Welch, *Removing Discriminatory Barriers: Basing Disparate Treatment Analysis on Motive Rather than Intent*, 60 S. CAL. L. REV. 733, 740 (1987).

have pointed out, motive and intent are not necessarily synonymous.<sup>115</sup> Motive can be read more broadly to include unconscious discrimination, including anything that influences a person to act, such as emotions or desires.<sup>116</sup> Nonetheless, courts have conflated the meanings of motive and intent such that the phrase “motive or intent” has come to refer only to conscious choices.<sup>117</sup> Thus, while most individual decision making probably belongs in a mixed-motive framework, as each decision a person makes comprises a complicated mix of motivations,<sup>118</sup> the mixed-motive framework will be no better than the pretext framework at addressing bias that occurs absent conscious intent.<sup>119</sup>

Except for masking, discriminatory data mining is by stipulation unintentional. Unintentional disparate treatment is not a problem that is new to data mining. A vast scholarly literature has developed regarding the law’s treatment of unconscious, implicit bias.<sup>120</sup> Such treatment can occur when an employer has internalized some racial stereotype and applies it or, without realizing it, monitors an employee more closely until the employer finds a violation.<sup>121</sup> The employee is clearly treated differently, but it is not intentional, and the employer is unaware of it. As Professor Samuel Bagenstos summarized, at this point, “it may be difficult, if not impossible, for a court to go back and reconstruct the numerous biased evaluations and perceptions that ultimately resulted in an adverse employment decision.”<sup>122</sup> Within the scholarly literature, there is “[s]urprising unanimity” that the law does not adequately address unconscious disparate treatment.<sup>123</sup>

115. Krieger, *supra* note 89, at 1243; Sullivan, *supra* note 114, at 915.

116. Krieger, *supra* note 89, at 1243; Sullivan, *supra* note 114, at 915 n.18 (quoting *Motive*, OXFORD ENGLISH DICTIONARY (1st ed. 1933)).

117. Sullivan, *supra* note 114, at 914–16, 916 n.20.

118. Amy L. Wax, *Discrimination as Accident*, 74 IND. L.J. 1129, 1149 & n.21 (1999); Krieger, *supra* note 89, at 1223. In fact, after the Supreme Court decided *Desert Palace*, many scholars thought that it *had* effectively overruled the *McDonnell-Douglas* framework, forcing all disparate treatment cases into a mixed-motive framework. *See, e.g.*, Sullivan, *supra* note 114, at 933–36 (discussing the then-emerging scholarly consensus). This has not played out so far, with courts and scholars split on the matter. *See, e.g.*, Kendall D. Isaac, *Is It “A” or Is It “The”?* *Deciphering the Motivating-Factor Standard in Employment Discrimination and Retaliation Cases*, 1 TEX. A&M L. REV. 55, 74 (2013) (“*McDonnell Douglas* has never been overruled and remains widely utilized.”); Barrett S. Moore, *Shifting the Burden: Genuine Disputes and Employment Discrimination Standards of Proof*, 35 U. ARK. LITTLE ROCK L. REV. 113, 123–29, 128 n.146 (2012) (noting a circuit split on the issue).

119. *See* Krieger, *supra* note 89, at 1182–83.

120. *See, e.g.*, Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CALIF. L. REV. 969, 978 n.45 (2006) (collecting sources); Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment*, 94 CALIF. L. REV. 997, 1003 n.21 (2006) (collecting sources).

121. This example can be ported directly to data mining as overrepresentation in data collection. *See supra* Part I.B.2.

122. Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 9 (2006).

123. Sullivan, *supra* note 114, at 1000. There is, however, no general agreement on whether the law should treat such discrimination as disparate treatment or disparate impact. *Compare* Krieger, *supra* note 89, at 1231 (explaining that because the bias causes employers to *treat* people differently, it

There are a few possible ways to analogize discriminatory data mining to unintentional disparate treatment in the traditional context, based on where one believes the “treatment” lies. Either the disparate treatment occurs at the decision to apply a predictive model that will treat members of a protected class differently, or it occurs when the disparate result of the model is used in the ultimate hiring decision. In the first scenario, the intent at issue is the decision to apply a predictive model with known disproportionate impact on protected classes. In the second, the disparate treatment occurs if, after the employer sees the disparate result, he proceeds anyway. If the employer continues *because* he liked the discrimination produced in either scenario, then intent is clear. If not, then this just devolves into a standard disparate impact scenario, with liability based on effect. Under disparate impact theory, deciding to follow through on a test with discriminatory effect does not suddenly render it disparate *treatment*.<sup>124</sup>

Another option is to imagine the *model* as the decision maker exhibiting implicit bias. That is, because of biases hidden to the predictive model such as nonrepresentative data or mislabeled examples, the model reaches a discriminatory result. This analogy turns every mechanism except proxy discrimination into the equivalent of implicit bias exhibited by individual decision makers. The effect of bias is one factor among the many different factors that go into the model-driven decision, just like in an individual’s adverse employment decision.<sup>125</sup> Would a more expansive definition of motive fix this scenario?

Because the doctrine focuses on *human* decision makers as discriminators, the answer is no. Even if disparate treatment doctrine could capture unintentional discrimination, it would only address such discrimination stemming from human bias. For example, the person who came up with the idea for Street Bump ultimately devised a system that suffers from reporting bias,<sup>126</sup> but it was not because he or she was implicitly employing some racial stereotype. Rather, it was simply inattentiveness to problems with the sampling frame. This is not to say that his or her own bias had nothing to do with it—the person likely owned a smartphone and thus did not think about the people who do not—but no one would say that it was even implicit bias against protected

---

should be considered a disparate treatment violation), *with Sullivan, supra* note 114, at 969–71 (arguing that the purpose of disparate impact is a catch-all provision to address those types of bias that disparate treatment cannot reach). This disagreement is important and even more pronounced in the case of data mining. *See infra* Part III. For now, we assume each case can be analyzed separately.

124. In fact, after *Ricci v. DeStefano*, 557 U.S. 557 (2009), deciding *not* to apply such a test after noticing the discriminatory effect may give rise to a disparate treatment claim in the other direction.

125. Bagenstos, *supra* note 122, at 9; Krieger, *supra* note 89, at 1185–86 (“Not only disparate treatment analysis, but the entire normative structure of Title VII’s injunction ‘not to discriminate,’ rests on the assumption that decisionmakers possess ‘transparency of mind’—that they are aware of the reasons why they are about to make, or have made, a particular employment decision.”).

126. *See supra* note 51 and accompanying text.

classes that motivated the decision, even under the expansive definition of the word “motive.”<sup>127</sup>

The only possible analogy relevant to disparate treatment, then, is to those data mining mechanisms of unintentional discrimination that reflect a real person’s bias—something like LinkedIn’s Talent Match recommendation engine, which relies on potentially prejudiced human assessments of employees.<sup>128</sup> As a general rule, an employer may not avoid disparate treatment liability by encoding third-party preferences as a rationale for a hiring decision.<sup>129</sup> But, once again, to be found liable under current doctrine, the employer would likely both have to know that this is the specific failure mechanism of the model and choose it based on this fact.

There is one other interesting question regarding disparate treatment doctrine: whether the intent standard includes knowledge. This is not a problem that arises often when a human is making a single employment determination. Assuming disparate treatment occurs in a given case, it is generally either intended or unconscious. What would it mean to have an employer *know* that he was treating an employee differently, but still take the action he had always planned to take without *intent* to treat the employee differently? It seems like an impossible line to draw.<sup>130</sup>

With data mining, though, unlike unconscious bias, it is possible to audit the resulting model and inform an employer that she will be treating individuals differently before she does so. If an employer *intends* to employ the model, but *knows* it will produce a disparate impact, does she intend to discriminate? This is a more realistic parsing of intent and knowledge than in the case of an individual, nonsystematic employment decision. Neither pretext nor motive exists here, and throughout civil and criminal law, “knowledge” and “intent” are considered distinct states of mind, so there would likely be no liability. On the other hand, courts may use knowledge of discrimination as evidence to find intent.<sup>131</sup> And while the statute’s language only covers intentional discrimination,<sup>132</sup> a broad definition of intent could include knowledge or

---

127. Of course, the very presumption of a design’s neutrality is itself a bias that may work against certain people. See Langdon Winner, *Do Artifacts Have Politics?*, 109 DAEDALUS 121, 125 (1980). But, as this is a second-order effect, we need not address it here.

128. See Woods, *supra* note 46.

129. See 29 C.F.R. § 1604.2(a)(1)(iii) (2015) (stating the EEOC’s position that “the preferences of coworkers, the employer, clients or customers” cannot be used to justify disparate treatment); see also *Fernandez v. Wynn Oil Co.*, 653 F.2d 1273, 1276–77 (9th Cir. 1981); *Diaz v. Pan Am. World Airways, Inc.*, 442 F.2d 385, 389 (5th Cir. 1971).

130. See Krieger, *supra* note 89, at 1185 (discussing disparate treatment’s “assumption of decisionmaker self-awareness”).

131. *Columbus Bd. of Educ. v. Penick*, 443 U.S. 449, 464 (1979) (“[A]ctions having foreseeable and anticipated disparate impact are relevant evidence to prove the ultimate fact, forbidden purpose.”); *Pers. Adm’r of Mass. v. Feeney*, 442 U.S. 256, 279 n.25 (1979) (“[W]hen the adverse consequences of a law upon an identifiable group are . . . inevitable . . . , a strong inference that the adverse effects were desired can reasonably be drawn.”).

132. 42 U.S.C. § 2000e-2(h) (2012).

substantial certainty of the result.<sup>133</sup> Because the situation has not come up often, the extent of the “intent” required is as yet unknown.<sup>134</sup>

In sum, aside from rational racism and masking (with some difficulties), disparate treatment doctrine does not appear to do much to regulate discriminatory data mining.

### B. Disparate Impact

Where there is no discriminatory intent, disparate impact doctrine should be better suited to finding liability for discrimination in data mining. In a disparate impact case, a plaintiff must show that a particular facially neutral employment practice causes a disparate impact with respect to a protected class.<sup>135</sup> If shown, the defendant-employer may “demonstrate that the challenged practice is job related for the position in question and consistent with business necessity.”<sup>136</sup> If the defendant makes a successful showing to that effect, the plaintiff may still win by showing that the employer could have used an “alternative employment practice” with less discriminatory results.<sup>137</sup>

The statute is unclear as to the required showing for essentially every single element of a disparate impact claim. First, it is unclear how much disparate impact is needed to make out a prima facie case.<sup>138</sup> The EEOC, charged with enforcing Title VII’s mandate, has created the so-called “four-fifths rule” as a presumption of adverse impact: “A selection rate for any race, sex, or ethnic group which is less than four-fifths . . . of the rate for the group

---

133. See Julia Kobick, Note, *Discriminatory Intent Reconsidered: Folk Concepts of Intentionality and Equal Protection Jurisprudence*, 45 HARV. C.R.-C.L. L. REV. 517, 551 (2010) (arguing that courts should regularly consider knowledge and foreseeability of disparate impact as an intended effect); cf. RESTATEMENT (SECOND) OF TORTS § 8A cmt. b (AM. LAW INST. 1965) (“Intent is not . . . limited to consequences which are desired. If the actor knows that the consequences are certain, or substantially certain, to result from his act, and still goes ahead, he is treated by the law as if he had in fact desired to produce the result.”).

134. Determining that a model is discriminatory is also like trying and failing to validate a test under disparate impact doctrine. See *infra* Part II.B. If a test fails validation, the employer using it would know that he is discriminating if he applies it, but that does not imply that he is subject to disparate treatment liability. Nonetheless, validation is part of the business necessity defense, and that defense is not available against disparate treatment claims. Thus, the analysis does not necessarily have the same result. 42 U.S.C. § 2000e-2(k)(2). One commentator has argued that including knowledge as a state of mind leading to disparate treatment liability would effectively collapse disparate impact and disparate treatment by conflating intent and effect. Jessie Allen, *A Possible Remedy for Unthinking Discrimination*, 61 BROOK. L. REV. 1299, 1314 (1995). But others still have noted that with respect to knowledge, a claim is still about the *treatment* of an individual, not the incidental disparate impact of a neutral policy. See Carin Ann Clauss, *Comparable Worth—The Theory, Its Legal Foundation, and the Feasibility of Implementation*, 20 U. MICH. J.L. REFORM 7, 62 (1986).

135. 42 U.S.C. § 2000e-2(k)(1)(A).

136. *Id.*

137. *Id.*

138. The statute does not define the requirement and Supreme Court has never addressed the issue. See, e.g., Sullivan, *supra* note 114, at 954 & n.153. For a brief discussion of the different approaches to establishing disparate impact, see Pamela L. Perry, *Two Faces of Disparate Impact Discrimination*, 59 FORDHAM L. REV. 523, 570–74 (1991).

with the highest rate will generally be regarded . . . as evidence of adverse impact.”<sup>139</sup> The Uniform Guidelines on Employment Selection Procedures (Guidelines) also state, however, that smaller differences can constitute adverse impact and greater differences may not, depending on circumstances. Thus, the four-fifths rule is truly just a guideline.<sup>140</sup> For the purposes of this Part, it is worthwhile to just assume that the discriminatory effects are prominent enough to establish disparate impact as an initial matter.<sup>141</sup>

The next step in the litigation is the “business necessity” defense. This defense is, in a very real sense, the crux of disparate impact analysis, weighing Title VII’s competing goals of limiting the effects of discrimination while allowing employers discretion to advance important business goals. *Griggs v. Duke Power Co.*<sup>142</sup>—the decision establishing the business necessity defense alongside disparate impact doctrine itself—articulated the defense in several different ways:

A challenged employment practice must be “shown to be related to job performance,” have a “manifest relationship to the employment in question,” be “demonstrably a reasonable measure of job performance,” bear some “relationship to job-performance ability,” and/or “must measure the person for the job and not the person in the abstract.”<sup>143</sup>

The Supreme Court was not clear on what, if any, difference existed between job-relatedness and business necessity, at one point seeming to use the terms interchangeably: “The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited.”<sup>144</sup> The focus of the Court was clearly on future job performance, and the term “job-related” has come to mean a practice that is predictive of job performance.<sup>145</sup> Because the definitions of job-relatedness and business necessity have never been clear, courts defer when applying the doctrine and finding the appropriate balance.<sup>146</sup>

Originally, the business necessity defense seemed to apply narrowly. In *Griggs*, Duke Power had instituted new hiring requirements including a high school diploma and success on a “general intelligence” test for previously

139. Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D) (2015) [hereinafter Guidelines].

140. *Id.*

141. We will return to this when discussing the need to grapple with substantive fairness. *See infra* Part III.B.

142. 401 U.S. 424 (1971).

143. Linda Lye, Comment, *Title VII’s Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense*, 19 BERKELEY J. EMP. & LAB. L. 315, 321 (1998) (footnotes omitted) (quoting *Griggs v. Duke Power Co.*, 401 U.S. 424, 431–36 (1971)).

144. *Griggs*, 401 U.S. at 431; *see also* Lye, *supra* note 143, at 320.

145. Lye, *supra* note 143, at 355 & n.206.

146. *Id.* at 319–20, 348–53; Amy L. Wax, *Disparate Impact Realism*, 53 WM. & MARY L. REV. 621, 633–34 (2011).



white-only divisions. Duke Power did not institute such requirements in divisions where it had previously hired black employees.<sup>147</sup> The Court ruled that the new requirements were not a business necessity because “employees who have not completed high school or taken the tests have continued to perform satisfactorily and make progress in departments for which the high school and test criteria are now used.”<sup>148</sup> Furthermore, the requirements were implemented without any study of their future effect.<sup>149</sup> The Court also rejected the argument that the requirements would improve the “overall quality of the workforce.”<sup>150</sup>

By 1979, the Court began treating business necessity as a much looser standard.<sup>151</sup> In *New York City Transit Authority v. Beazer*,<sup>152</sup> the transit authority had implemented a rule barring drug users from employment, including current users of methadone, otherwise known as *recovering* heroin addicts. In dicta, the Court stated that a “narcotics rule,” which “significantly serves” the “legitimate employment goals of safety and efficiency,” was “assuredly” job related.<sup>153</sup> This was the entire analysis of the business necessity defense in the case. Moreover, the rationale was acceptable as applied to the entire transit authority, even where only 25 percent of the jobs were labeled as “safety sensitive.”<sup>154</sup> Ten years later, the Court made the business necessity doctrine even more defendant-friendly in *Wards Cove Packing Co. v. Atonio*.<sup>155</sup> After *Wards Cove*, the business necessity defense required a court to engage in “a reasoned review of the employer’s justification for his use of the challenged practice. . . . [T]here is no requirement that the challenged practice be ‘essential’ or ‘indispensable’ to the employer’s business for it to pass muster . . . .”<sup>156</sup> The Court also reallocated the burden to plaintiffs to prove that business necessity was lacking and even referred to the defense as a “business justification” rather than a business necessity.<sup>157</sup> The *Wards Cove* Court went so far that Congress directly addressed the decision in the Civil Rights Act of 1991 (1991 Act), which codified disparate impact and reset the standards to the day before *Wards Cove* was decided.<sup>158</sup>

Because the substantive standards for job-relatedness or business necessity were uncertain before *Wards Cove*, however, the confusion persisted

---

147. *Griggs*, 401 U.S. at 427–28.

148. *Id.* at 431–32.

149. *Id.* at 432.

150. *Id.* at 431.

151. See Nicole J. DeSario, *Reconceptualizing Meritocracy: The Decline of Disparate Impact Discrimination Law*, 38 HARV. C.R.-C.L. L. REV. 479, 495–96 (2003); Lye, *supra* note 143, at 328.

152. 440 U.S. 568 (1979).

153. *Id.* at 587 & n.31.

154. *Id.*

155. 490 U.S. 642 (1989).

156. *Id.* at 659.

157. *Id.*

158. 42 U.S.C. § 2000e-2(k)(1)(C) (2012).

even after the 1991 Act was passed.<sup>159</sup> At the time, both sides—civil rights groups and the Bush administration, proponents of a rigorous and more lenient business necessity defense respectively—declared victory.<sup>160</sup>

Since then, courts have recognized that business necessity lies somewhere in the middle of two extremes.<sup>161</sup> Some courts require that the hiring criteria bear a “manifest relationship”<sup>162</sup> to the employment in question or that they be “significantly correlated” to job performance.<sup>163</sup> The Third Circuit was briefly an outlier, holding “that hiring criteria must effectively measure the ‘minimum qualifications for successful performance of the job’” in order to meet the strict business necessity standard.<sup>164</sup> This tougher standard would, as a practical matter, ban general aptitude tests with any disparate impact because a particular cutoff score cannot be shown to distinguish between those able and completely unable to do the work.<sup>165</sup> For example, other unmeasured skills and abilities could theoretically compensate for the lower score on an aptitude test, rendering a certain minimum score not “necessary” if it does not measure minimum qualifications.<sup>166</sup> In a subsequent case, however, the Third Circuit recognized that Title VII does not require an employer to choose someone “less qualified” (as opposed to unqualified) in the name of nondiscrimination and noted that aptitude tests can be legitimate hiring tools if they accurately measure a person’s qualifications.<sup>167</sup> The court concluded:

---

159. Legislative history was no help either. The sole piece of legislative history is an “interpretive memorandum” that specifies that the standards were to revert to before *Wards Cove*, coupled with an explicit instruction in the Act to ignore any other legislative history regarding business necessity. Susan S. Grover, *The Business Necessity Defense in Disparate Impact Discrimination Cases*, 30 GA. L. REV. 387, 392–93 (1996).

160. Andrew C. Spiropoulos, *Defining the Business Necessity Defense to the Disparate Impact Cause of Action: Finding the Golden Mean*, 74 N.C. L. REV. 1479, 1484 (1996).

161. Though courts generally state the standard to reflect this middle position, the Supreme Court’s latest word on disparate impact—in which the Court reaffirmed the doctrine generally and held that it applied in the Fair Housing Act—included the decidedly defendant-friendly observation that “private policies are not contrary to the disparate-impact requirement unless they are ‘artificial, arbitrary, and unnecessary barriers.’” *Tex. Dep’t of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc.*, 135 S. Ct. 2507, 2512 (2015) (quoting *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971)).

162. See, e.g., *Gallagher v. Magner*, 619 F.3d 823, 834 (8th Cir. 2010); *Anderson v. Westinghouse Savannah River Co.*, 406 F.3d 248, 265 (4th Cir. 2005).

163. *Gulino v. N.Y. State Educ. Dep’t*, 460 F.3d 361, 383 (2d Cir. 2006) (noting that hiring criteria are “significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated” (quoting *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 431 (1975))).

164. *El v. Se. Pa. Transp. Auth.*, 479 F.3d 232, 242 (3d Cir. 2007) (quoting *Lanning v. Se. Pa. Transp. Auth.*, 181 F.3d 478, 481 (3d Cir. 1999)).

165. Michael T. Kirkpatrick, *Employment Testing: Trends and Tactics*, 10 EMP. RTS. & EMP. POL’Y J. 623, 633 (2006).

166. *Id.* Note, though, that this is similar to arguing that there is a less discriminatory alternative employment practice. This argument, then, would place the burden of the alternative employment practice prong on the defendant, contravening the burden-shifting scheme in the statute. See *infra* notes 170–74 and accompanying text.

167. *El*, 479 F.3d at 242.

Putting these standards together, then, we require that employers show that a discriminatory hiring policy accurately—but not perfectly—ascertains an applicant's ability to perform successfully the job in question. In addition, Title VII allows the employer to hire the applicant most likely to perform the job successfully over others less likely to do so.<sup>168</sup>

Thus, all circuits seem to accept varying levels of job-relatedness rather than strict business necessity.<sup>169</sup>

The last piece of the disparate impact test is the “alternative employment practice” prong. Shortly after *Griggs*, the Supreme Court decided *Albemarle Paper Co. v. Moody*, holding in part that “[i]f an employer does then meet the burden of proving that its tests are ‘job related,’ it remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer’s legitimate interest in ‘efficient and trustworthy workmanship.’”<sup>170</sup> This burden-shifting scheme was codified in the 1991 Act as the “alternative employment practice” requirement.<sup>171</sup> Congress did not define the phrase, and its substantive meaning

168. *Id.*

169. Interestingly, it seems that many courts read identical business necessity language in the Americans with Disabilities Act to refer to a minimum qualification standard. *See, e.g., Sullivan v. River Valley Sch. Dist.*, 197 F.3d 804, 811 (6th Cir. 1999) (“[T]here must be significant evidence that could cause a reasonable person to inquire as to whether an employee is still capable of performing his job. An employee’s behavior cannot be merely annoying or inefficient to justify an examination; rather, there must be genuine reason to doubt whether that employee can ‘perform job-related functions.’” (quoting 42 U.S.C. § 12112(d)(4)(B))). Presumably, this is because disability, when compared to race or sex, more immediately raises questions regarding a person’s ability to perform a job. Ironically, however, this means that disparate impact will be *more* tolerated where it is less likely to be obviously justified. Christine Jolls has in fact argued that disparate impact is, to a degree, functionally equivalent to accommodations law. Jolls, *supra* note 90, at 652.

170. 422 U.S. 405, 425 (1975) (quoting *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 801 (1973)).

171. 42 U.S.C. § 2000e-2(k)(1)(A) (2012). The “alternative employment practice” test has not always been treated as a separate step. *See, e.g., Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 659 (1989) (treating the alternative employment practice test as part of the “business justification” phase); *Dothard v. Rawlinson*, 433 U.S. 321, 332 (1977) (treating the alternative employment practice test as a narrow tailoring requirement for the business necessity defense). The *Albemarle* Court, though creating a surrebuttal and thus empowering plaintiffs, seemed to regard the purpose of disparate impact as merely smoking out pretexts for intentional discrimination. 422 U.S. at 425; *see also Primus, supra* note 98, at 537. If the *Albemarle* Court’s approach is correct, treating the alternative employment practice requirement as a narrow tailoring requirement does make sense, much as the narrow tailoring requirement of strict scrutiny in equal protection serves the function of smoking out invidious purpose. *City of Richmond v. J.A. Croson Co.*, 488 U.S. 469, 493 (1989); *Rubinfeld, supra* note 99, at 428.

Every circuit to address the question, though, has held that the 1991 Act returned the doctrine to the *Albemarle* burden-shifting scheme. *Jones v. City of Boston*, 752 F.3d 38, 54 (1st Cir. 2014); *Howe v. City of Akron*, 723 F.3d 651, 658 (6th Cir. 2013); *Tabor v. Hilti, Inc.*, 703 F.3d 1206, 1220 (10th Cir. 2013); *Puffer v. Allstate Ins. Co.*, 675 F.3d 709, 717 (7th Cir. 2012); *Gallagher v. Magner*, 619 F.3d 823, 833 (8th Cir. 2010); *Gulino v. N.Y. State Educ. Dep’t*, 460 F.3d 361, 382 (2d Cir. 2006); *Int’l Bhd. of Elec. Workers Local Unions Nos. 605 & 985 v. Miss. Power & Light Co.*, 442 F.3d 313, 318 (5th Cir. 2006); *Anderson v. Westinghouse Savannah River Co.*, 406 F.3d 248, 277

remains uncertain. *Wards Cove* was the first case to use the specific phrase, so Congress's instruction to reset the law to the pre-*Wards Cove* standard is particularly perplexing.<sup>172</sup> The best interpretation is most likely *Albemarle*'s reference to "other tests or selection devices, without a similarly undesirable racial effect."<sup>173</sup> But this interpretation is slightly odd because in *Albemarle*, business necessity was still somewhat strict, and it is hard to imagine a business practice that is "necessary" while there exists a less discriminatory alternative that is just as effective.<sup>174</sup> If business necessity or job-relatedness is a less stringent requirement, though, then the presence of the alternative employment practice requirement does at least give it some teeth.

Now return to data mining. For now, assume a court does not apply the strict business necessity standard but has some variation of "job related" in mind (as all federal appellate courts do today).<sup>175</sup> The threshold issue is clearly whether the sought-after trait—the target variable—is job related, regardless of the machinery used to predict it. If the target variable is not sufficiently job related, a business necessity defense would fail, regardless of the fact that the decision was made by algorithm. Thus, disparate impact liability can be found for improper care in target variable definition. For example, it would be difficult for an employer to justify an adverse determination based on the appearance of an advertisement suggesting a criminal record alongside the search results for a candidate's name. Sweeney found such a search to have a disparate impact,<sup>176</sup> and the EEOC and several federal courts have interpreted Title VII to prohibit discrimination on the sole basis of criminal record, unless there is a specific reason the particular conviction is related to the job.<sup>177</sup> This

---

(4th Cir. 2005); Ass'n of Mexican-Am. Educators v. California, 231 F.3d 572, 584 (9th Cir. 2000); EEOC v. Joe's Stone Crab, Inc., 220 F.3d 1263, 1275 (11th Cir. 2000); Lanning v. Se. Pa. Transp. Auth., 181 F.3d 478, 485 (3d Cir. 1999). The D.C. Circuit has not explicitly observed that a burden-shifting framework exists.

172. Sullivan, *supra* note 114, at 964; Michael J. Zimmer, *Individual Disparate Impact Law: On the Plain Meaning of the 1991 Civil Rights Act*, 30 LOY. U. CHI. L.J. 473, 485 (1999).

173. *Albemarle*, 422 U.S. at 425; accord, e.g., *Jones*, 752 F.3d at 53 (citing *Albemarle* to find meaning in the 1991 Act's text); *Allen v. City of Chicago*, 351 F.3d 306, 312 (7th Cir. 2003) (same, but with a "see also" signal).

174. William R. Corbett, *Fixing Employment Discrimination Law*, 62 SMU L. REV. 81, 92 (2009).

175. The difference would be whether mining for a single job-related trait, rather than a holistic ranking of "good employees," is permissible at all. See *infra* text accompanying notes 197–99.

176. Sweeney, *supra* note 41, at 51.

177. See *El v. Se. Pa. Transp. Auth.*, 479 F.3d 232, 243 (3d Cir. 2007) (finding that though the criminal record policy had a disparate impact, it satisfied business necessity in that case); *Green v. Mo. Pac. R.R.*, 523 F.2d 1290, 1298 (8th Cir. 1975); *McCain v. United States*, No. 2:14-cv-92, 2015 WL 1221257, at \*17 (D. Vt. Mar. 17, 2015); EQUAL EMP'T OPPORTUNITY COMM'N, CONSIDERATION OF ARREST AND CONVICTION RECORDS IN EMPLOYMENT DECISIONS UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964 (2012), [http://www.eeoc.gov/laws/guidance/upload/arrest\\_conviction.pdf](http://www.eeoc.gov/laws/guidance/upload/arrest_conviction.pdf) [<https://perma.cc/JY47-2HVT>]; see also *Univ. of Tex. Sw. Med. Ctr. v. Nassar*, 133 S. Ct. 2517, 2540 (2013) ("The position set out in the EEOC's guidance and compliance manual merits respect."); Michael Connett, Comment, *Employer Discrimination Against Individuals with a Criminal Record: The Unfulfilled Role of State Fair Employment Agencies*, 83 TEMP. L. REV. 1007, 1017 & nn.82–83

is true independent of the fact that the disparity is an artifact of third-party bias; all that matters is whether the target variable is job related. In the end, though, because determining that a business practice is not job related actually requires a normative determination that it is instead discriminatory, courts tend to accept most common business practices for which an employer has a plausible story.<sup>178</sup>

Once a target variable is established as job related, the first question is whether the model is predictive of that trait. The nature of data mining suggests that this will be the case. Data mining is designed entirely to predict future outcomes, and, if seeking a job-related trait, future job performance. One commentator lamented that “[f]ederal case law has shifted from a prospective view of meritocracy to a retrospective view, thereby weakening disparate impact law.”<sup>179</sup> The author meant that, in *Griggs*, the Court recognized that education and other external factors were unequal and therefore discounted a measure of meritocracy that looked to past achievements, in favor of comparing the likelihood of future ones. But by the time the Court had decided *Wards Cove*, it had shifted to a model of retrospective meritocracy that presumed the legitimacy of past credentials, thus upholding the status quo.<sup>180</sup> While data mining must take the past—represented by the training data—as given, it generates predictions about workplace success that are much more accurate than predictions based on those past credentials that disparate impact doctrine has come to accept.<sup>181</sup> In a hypothetical perfect case of data mining, the available information would be rich enough that reliance on the past information would fully predict future performance. Thus, robust data mining would likely satisfy even the *Griggs* Court’s standard that the models are looking toward future job performance, not merely past credentials.

The second question asks whether the model adequately predicts what it is supposed to predict. In the traditional context, this question arises in the case of general aptitude tests that might end up measuring unrelated elements of cultural awareness rather than intelligence.<sup>182</sup> This is where the different data

---

(2011) (citing EQUAL EMP’T OPPORTUNITY COMM’N, POLICY STATEMENT ON THE ISSUE OF CONVICTION RECORDS UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964 (1987), <http://www.eeoc.gov/policy/docs/convict1.html> [<https://perma.cc/PY24-V8V7>]). *But see, e.g.*, *Manley v. Invesco*, 555 Fed. App’x 344, 348 (5th Cir. 2014) (per curiam) (“Persons with criminal records are not a protected class under Title VII.”).

178. Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 753 (2006).

179. DeSario, *supra* note 151, at 481.

180. *Id.* at 493; *see also infra* Conclusion.

181. *See* Don Peck, *They’re Watching You at Work*, ATLANTIC (Nov. 20, 2013), <http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681> [<https://perma.cc/JFP8-CZKC>] (discussing Google’s choice to abandon traditional hiring metrics because they are not good predictors of performance).

182. *See, e.g.*, *Griggs v. Duke Power Co.*, 420 F.2d 1225, 1239 n.6 (4th Cir. 1970), *rev’d*, 401 U.S. 424 (1971) (“Since for generations blacks have been afforded inadequate educational opportunities and have been culturally segregated from white society, it is no more surprising that their

mining mechanisms for discriminatory effects matter. Part I posited that proxy discrimination optimizes correctly. So if it evidences a disparate impact, it reflects unequal distribution of relevant traits in the real world. Therefore, proxy discrimination will be as good a job predictor as possible given the current shape of society. Models trained on biased samples and mislabeled examples, on the other hand, will result in correspondingly skewed assessments rather than reflect real-world disparities. The same effect may be present in models that rely on insufficiently rich or insufficiently granular datasets: by designation they do not reflect reality. These models might or might not be considered job related, depending on whether the errors distort the outcomes enough that the models are no longer good predictors of job performance.

The Guidelines have set forth validation procedures intended to create a job-relatedness standard. Quantifiable tests that have a disparate impact must be validated according to the procedures in the Guidelines if possible; otherwise, a presumption arises that they are not job related.<sup>183</sup> Under the Guidelines, a showing of validity takes one of three forms: criterion-related, content, or construct.<sup>184</sup> Criterion-related validity “consist[s] of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance.”<sup>185</sup> The “relationship between performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance. . . .”<sup>186</sup> Content validity refers to testing skills or abilities that generally are or have been learned on the job, though not those that could be acquired in a “brief orientation.”<sup>187</sup> Construct validity refers to a test designed to measure some innate human trait such as honesty. A user of a construct “should show by empirical evidence that the selection procedure is validly related to the construct and that the construct is validly related to the performance of critical or important work behavior(s).”<sup>188</sup>

As a statistical predictive measure, a data mining model could be validated by either criterion-related or construct validity, depending on the trait being sought. Either way, there must be statistical significance showing that the result of the model correlates to the trait (which was already determined to be an important element of job performance). This is an exceedingly low bar for data mining because data mining’s predictions necessarily rest on demonstrated

---

performance on ‘intelligence’ tests is significantly different than whites’ than it is that fewer blacks have high school diplomas.”).

183. 29 C.F.R. §§ 1607.3, 1607.5 (2015). The Guidelines also cite two categories of practices that are unsuitable for validation: informal, unscored practices and technical infeasibility. *Id.* § 1607.6(B). For the latter case, the Guidelines state that the selection procedure still should be justified somehow or another option should be chosen.

184. *Id.* § 1607.5(B).

185. *Id.*

186. *Id.* § 1607.14(B)(5).

187. *Id.* §§ 1607.5(F), 1607.14(C).

188. *Id.* § 1607.14(D)(3).

statistical relationships. Data mining will likely only be used if it is actually predictive of *something*, so the business necessity defense solely comes down to whether the trait sought is important enough to job performance to justify its use in any context.

Even assuming the Guidelines' validation requirement is a hurdle for data mining, some courts ignore the Guidelines' recommendation that an unvalidated procedure be rejected, preferring to rely on "common sense" or finding a "manifest relationship" between the criteria and successful job performance.<sup>189</sup> Moreover, it is possible that the Supreme Court inadvertently overruled the Guidelines in 2009. In *Ricci v. Destefano*, a case that will be discussed in greater detail in Part III.B, the Court found no genuine dispute that the tests at issue met the job-related and business necessity standards<sup>190</sup> despite not having been validated under the Guidelines and despite the employer *actively denying* that they could be validated.<sup>191</sup> While the business necessity defense was not directly at issue in *Ricci*, "[o]n the spectrum between heavier and lighter burdens of justification, the Court came down decidedly in favor of a lighter burden."<sup>192</sup>

Thus, there is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense. Models trained on biased samples, mislabeled examples, and limited features, however, might trigger liability under the alternative employment practice prong. If a plaintiff can show that an alternative, less discriminatory practice that accomplishes the same goals exists and that the employer "refuses" to use it, the employer can be found liable. In this case, a plaintiff could argue that the obvious alternative employment practice would be to fix the problems with the models.

Fixing the models, however, is not a trivial task. For example, in the LinkedIn hypothetical, where the demonstrated interest in different kinds of employees reflects employers' prejudice, LinkedIn is the party that determines the algorithm by which the discrimination occurs (in this case, based on reacting to third-party preferences). If an employer were to act on the recommendations suggested by the LinkedIn recommendation engine, there

---

189. Wax, *supra* note 146, at 633–34.

190. David A. Drachler, *Assessing the Practical Repercussions of Ricci*, AM. CONST. SOC'Y BLOG (July 27, 2009), <http://www.acslaw.org/node/13829> [<https://perma.cc/AH9G-B3GN>] (observing that the Court in *Ricci v. DeStefano* found no genuine dispute that the unvalidated tests at issue met the job-related and business necessity standards despite the Guidelines creating a presumption of invalidity for unvalidated tests that are discriminatory).

191. New Haven's primary argument was that it had to withdraw the tests or it would have faced Title VII liability. See Mark S. Brodin, *Ricci v. DeStefano: The New Haven Firefighters Case & the Triumph of White Privilege*, 20 S. CAL. REV. L. & SOC. JUST. 161, 178 n.128 (2011) ("New Haven forcefully argued throughout the litigation that the exams were 'flawed' and may not have identified the most qualified candidates for the supervisory positions.")

192. George Rutherglen, *Ricci v. Destefano: Affirmative Action and the Lessons of Adversity*, 2009 SUP. CT. REV. 83, 107.

would not be much he could do to make it less reflective of third-party prejudice, aside from calling LinkedIn and asking nicely. Thus, it could not really be said that the employer “refuses” to use an alternative employment practice. The employer could either use the third-party tool or not. Similarly, it might be possible to fix an app like Street Bump that suffers from reporting bias, but the employer would need access to the raw input data in order to do so.<sup>193</sup> In the case of insufficiently rich or granular features, the employer would need to collect more data in order to make the model more discerning. But collecting more data can be time consuming and costly,<sup>194</sup> if not impossible for legal or technical reasons.

Moreover, the under- and overrepresentation of members of protected classes in data is not always evident, nor is the mechanism by which such under- or overrepresentation occurs. The idea that the representation of different social groups in the dataset can be brought into proportions that better match those in the real world presumes that analysts have some independent mechanism for determining these proportions. Thus, there are several hurdles to finding disparate impact liability for models employing data that under- or overrepresents members of protected classes. The plaintiff must prove that the employer created or has access to the model, can discover that there is discriminatory effect, and can discover the particular mechanism by which that effect operates. The same can be said for models with insufficiently rich feature sets. Clearly there are times when more features would improve an otherwise discriminatory outcome. But it is, almost by definition, hard to know which features are going to make the model more or less discriminatory. Indeed, it is often impossible to know which features are missing because data miners do not operate with causal relationships in mind. So while theoretically a less discriminatory alternative would almost always exist, proving it would be difficult.

There is yet another hurdle. Neither Congress nor courts have specified what it means for an employer to “refuse” to adopt the less discriminatory procedure. Scholars have suggested that perhaps the employer cannot be held liable until it has considered the alternative and rejected it.<sup>195</sup> Thus, if the employer has run an expensive data collection and analysis operation without ever being made aware of its any discriminatory tendencies, and the employer cannot afford to re-run the entire operation, is the employer “refusing” to use a less discriminatory alternative, or does one simply not exist? How much would the error correction have to cost an employer before it is not seen as a refusal to use the procedure?<sup>196</sup> Should the statute actually be interpreted to mean that an

---

193. See *infra* Part III.B.1.

194. See generally Dalessandro, Perlich & Raeder, *supra* note 68.

195. Sullivan, *supra* note 114, at 964; Zimmer, *supra* note 172, at 505–06.

196. For a discussion of courts using cost as a rationale here, see Ernest F. Lidge III, *Financial Costs as a Defense to an Employment Discrimination Claim*, 58 ARK. L. REV. 1, 32–37 (2005).



employer “unreasonably refuses” to use an alternative employment practice? These are all difficult questions, but suffice it to say, the prospect of winning a data mining discrimination case on alternative employment practice grounds seems slim.

The third and final consideration regarding disparate impact liability for data mining is whether a court or Congress might reinvigorate strict business necessity.<sup>197</sup> In that case, things look a little better for plaintiffs bringing disparate impact claims. Where an employer models job tenure,<sup>198</sup> for example, a court may be inclined to hold that it is job related because the model is a “legitimate, non-discriminatory business objective.”<sup>199</sup> But it is clearly not necessary to the job. The same reasoning applies to mining for any single trait that is job related—the practice of data mining is not focused on discovering make-or-break skills. Unless the employer can show that below the cut score, employees cannot do the work, then the strict business necessity defense will fail. Thus, disparate impact that occurs as an artifact of the problem-specification stage can potentially be addressed by strict business necessity.

This reasoning is undermined, though, where employers do not mine for a single trait, but automate their decision process by modeling job performance on a holistic measure of what makes good employees. If employers determine traits of a good employee by simple ratings, and use data mining to appropriately divine good employees’ characteristics among several different variables, then the argument that the model does not account for certain skills that could compensate for the employee’s failings loses its force. Taken to an extreme, an 8,000-feature holistic determination of a “good employee” would still not be strictly “necessary.” Holding a business to such a standard, however, would simply be forbidding that business from ranking candidates if any disparate impact results. Thus, while the strict business necessity defense could prevent myopic employers from creating disparate impacts by their choice of target variable, it would still not address forms of data mining that model general job performance rather than predict specific traits.

Disparate impact doctrine was created to address unintentional discrimination. But it strikes a delicate balance between allowing businesses the leeway to make legitimate business judgments and preventing “artificial, arbitrary, and unnecessary” discrimination.<sup>200</sup> Successful data mining operations will often both predict future job performance and have some

---

197. This would likely require Congressional action because strict business necessity essentially transfers the burden to prove a lack of an alternative employment practice to the defense. By implication, if a practice is “necessary,” there cannot be alternatives. The statute, as it reads now, clearly states that the plaintiff has the burden for that prong. 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012).

198. This is an increasingly common practice in low-wage, high-turnover jobs. *See* Peck, *supra* note 181.

199. *Equal Emp’t Opportunity Comm’n v. Joe’s Stone Crab, Inc.*, 220 F.3d 1263, 1275 (11th Cir. 2000); *see also* *Gallagher v. Magner*, 619 F.3d 823, 834 (8th Cir. 2010).

200. *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971).

disparate impact. Unless the plaintiff can find an alternative employment practice to realistically point to, a tie goes to the employer.

### C. Masking and Problems of Proof

Masking poses separate problems for finding Title VII liability. As discussed earlier, there is no theoretical problem with finding liability for masking.<sup>201</sup> It is a disparate treatment violation as clear as any. But like traditional forms of intentional discrimination, it suffers from difficulties of proof. While finding intent from stray remarks or other circumstantial evidence is challenging in any scenario, masking presents additional complications for detection.

Data mining allows employers who wish to discriminate on the basis of a protected class to disclaim any knowledge of the protected class in the first instance while simultaneously inferring such details from the data. An employer may want to discriminate by using proxies for protected classes, such as in the case of redlining.<sup>202</sup> Due to housing segregation, neighborhood is a good proxy for race and can be used to redline candidates without reference to race.<sup>203</sup> This is a relatively unsophisticated example, however. It is possible that some combination of musical tastes,<sup>204</sup> stored “likes” on Facebook,<sup>205</sup> and network of friends<sup>206</sup> will reliably predict membership in protected classes. An employer can use these traits to discriminate by setting up future models to sort by these items and then disclaim any knowledge of such proxy manipulation.

More generally, as discussed in Part I, any of the mechanisms by which unintentional discrimination can occur can also be employed intentionally. The example described above is intentional discrimination by proxy, but it is also possible to intentionally bias the data collection process, purposefully mislabel examples, or deliberately use an insufficiently rich set of features,<sup>207</sup> though some of these would probably require a great deal of sophistication. These methods of intentional discrimination will look, for all intents and purposes, identical to the unintentional discrimination that can result from data mining. Therefore, detecting discrimination in the first instance will require the same techniques as detecting unintentional discrimination, namely a disparate impact analysis. Further, assuming there is no circumstantial evidence like an employer’s stray remarks with which to prove intent, a plaintiff might attempt

---

201. See *supra* text accompanying notes 106–07.

202. See *supra* Part I.E.

203. See MASSEY & DENTON, *supra* note 73, at 51–52.

204. Croll, *supra* note 88.

205. Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. NAT’L ACAD. SCI. 5802 (2013).

206. Carter Jernigan & Behram F.T. Mistree, *Gaydar: Facebook Friendships Expose Sexual Orientation*, FIRST MONDAY (Oct. 5, 2009), <http://firstmonday.org/article/view/2611/2302> [<https://perma.cc/G36G-S26X>].

207. See Dwork et al., *supra* note 81, app. at 226 (“Catalog of Evils”).

to prove intent by demonstrating that the employer is using less representative data, poorer examples, or fewer and less granular features than he might otherwise use were he interested in the best possible candidate. That is, one could show that the neutral employment practice is a pretext by demonstrating that there is a more predictive alternative.

This looks like disparate impact analysis. A plaintiff proving masked intentional discrimination asks the same question as in the “alternative employment practice” prong: whether there were more relevant measures the employer could have used.<sup>208</sup> But the business necessity defense is not available in a disparate treatment case,<sup>209</sup> so alternative employment practice is not the appropriate analysis. Scholars have noted, though, that the line between disparate treatment and disparate impact in traditional Title VII cases is not always clear,<sup>210</sup> and sometimes employer actions can be legitimately categorized as either or both.<sup>211</sup> As Professor George Rutherglen has pointed out, “Concrete issues of proof, more than any abstract theory, reveal the fundamental similarity between claims of intentional discrimination and those of disparate impact. The evidence submitted to prove one kind of claim invariably can be used to support the other.”<sup>212</sup> Rutherglen’s point is exactly what must happen in the data mining context: disparate treatment and disparate impact become essentially the same thing from an evidentiary perspective.

To the extent that disparate impact and treatment are, in reality, different theories, they are often confused for each other. Plaintiffs will raise both types of claims as a catch-all because they cannot be sure on which theory they might win, so both theories will be in play in a given case.<sup>213</sup> As a result, courts often seek evidence of state of mind in disparate impact cases<sup>214</sup> and objective, statistical evidence in disparate treatment cases.<sup>215</sup> Assuming the two theories are not functionally the same, using the same evidence for disparate treatment and disparate impact will only lead to more confusion and, as a result, more uncertainty within the courts. Thus, despite its clear nature as a theoretical violation, it is less clear that a plaintiff will be able to win a masking disparate treatment case.

A final point is that traditionally, employers who do *not* want to discriminate go to great lengths to avoid raising the prospect that they have

---

208. Cf. *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975) (creating an alternative employment practice prong for the purpose of rooting out pretext).

209. 42 U.S.C. § 2000e-2(k)(2) (2012).

210. George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 *FORDHAM L. REV.* 2313, 2313 (2006); Stacy E. Seicshnaydre, *Is the Road to Disparate Impact Paved with Good Intentions?: Stuck on State of Mind in Antidiscrimination Law*, 42 *WAKE FOREST L. REV.* 1141, 1142–43 (2007).

211. Rutherglen, *supra* note 210, at 2320–21.

212. *Id.* at 2320.

213. Seicshnaydre, *supra* note 210, at 1147–48.

214. *Id.* at 1153–63.

215. Rutherglen, *supra* note 210, at 2321–22.

violated the law. Thus they tend to avoid collecting information about attributes that reveal an individual's membership in a protected class. Employers even pay third parties to collect relatively easy-to-find information on job applicants, such as professional honors and awards, as well as compromising photos, videos, or membership in online groups, so that the third party can send back a version of the report that "remove[s] references to a person's religion, race, marital status, disability and other information protected under federal employment laws."<sup>216</sup> This allows employers to honestly disclaim any knowledge of the protected information. Nonetheless, if an employer seeks to discriminate according to protected classes, she would be able to infer class membership from the data. Thus, employers' old defense to suspicion of discrimination—that they did not even see the information—is no longer adequate to separate would-be intentional discriminators from employers that do not intend to discriminate.

### III.

#### THE DIFFICULTY FOR REFORMS

While each of the mechanisms for discrimination in data mining presents difficulties for Title VII as currently written, there are also certain obstacles to reforming Title VII to address the resulting problems. Computer scientists and others are working on technical remedies,<sup>217</sup> so to say that there are problems with legal remedies does not suggest that the problems with discrimination in data mining cannot be solved at all. Nonetheless, this Part focuses on the legal aspects. As it illustrates, even assuming that the political will to reform Title VII exists, potential legal solutions are not straightforward.

This Part discusses two types of difficulties with reforming Title VII. First, there are issues internal to the data mining process that make legal reform difficult. For example, the subjectivity in defining a "good employee" is unavoidable, but, at the same time, some answers are clearly less discriminatory than others.<sup>218</sup> How does one draw that line? Can employers gain access to the additional data necessary to correct for collection bias? How much will it cost them to find it? How do we identify the "correct" baseline historical data to avoid reproducing past prejudice or the "correct" level of detail and granularity in a dataset? Before laws can be reformed, policy-level answers to these basic technical, philosophical, and economic questions need to be addressed at least to some degree.

---

216. Jennifer Preston, *Social Media History Becomes a New Job Hurdle*, N.Y. TIMES (July 20, 2011), <http://www.nytimes.com/2011/07/21/technology/social-media-history-becomes-a-new-job-hurdle.html> [https://perma.cc/NZ8U-M296].

217. For a list of the wide-ranging research underway in computer science, see generally Resources, FAT ML, <http://www.fatml.org/resources.html> [https://perma.cc/T2QW-ARHX].

218. See *supra* Part I.A.

Second, reform will face political and constitutional constraints external to the logic of data mining that will affect how Title VII can be permissibly reformed to address it. Not all of the mechanisms for discrimination seem to be amenable to procedural remedies. If that holds true, only after-the-fact reweighting of results may be able to compensate for the discriminatory outcomes. This is not a matter of missing legislation; it is a matter of practical reality. Unfortunately, while in many cases no procedural remedy will be sufficient, any attempt to design a legislative or judicial remedy premised on reallocation of employment outcomes will not survive long in the current political or constitutional climate, as it raises the specter of affirmative action. Politically, anything that even hints at affirmative action is a nonstarter today, and to the extent that it is permissible to enact such policies, their future constitutionality is in doubt.<sup>219</sup>

### A. *Internal Difficulties*

#### 1. *Defining the Target Variable*

Settling on a target variable is a necessarily subjective exercise.<sup>220</sup> Disputes over the superiority of competing definitions are often insoluble because the target variables are themselves incommensurable. There are, of course, easier cases, where prejudice or carelessness leads to definitions that subject members of protected classes to avoidably high rates of adverse determinations. But most cases are likely to involve genuine business disagreements over ideal definitions, with each having a potentially greater or lesser impact on protected classes. There is no stable ground upon which to judge the relative merits of definitions because they often reflect competing ideas about the very nature of the problem at issue.<sup>221</sup> As Professor Oscar Gandy has argued, “[C]ertain kind[s] of biases are inherent in the selection of the goals or objective functions that automated systems will [be] designed to support.”<sup>222</sup> There is no escape from this situation; a target variable *must* reflect judgments about what really is the problem at issue in making hiring decisions. For certain employers, it might be rather obvious that the problem is one of reducing the administrative costs associated with turnover and training; for others, it might be improving sales; for still others, it might be increasing

---

219. See Lyle Denniston, *Argument Analysis: Now, Three Options on College Affirmative Action*, SCOTUSBLOG (Dec. 9, 2015, 2:47 PM), <http://www.scotusblog.com/2015/12/argument-analysis-now-three-options-on-college-affirmative-action> [<https://perma.cc/XF75-N82F>] (analysis of oral argument in *Fisher v. Univ. of Tex.*, 758 F.3d 633 (5th Cir. 2014), *cert. granted*, 135 S. Ct. 2888, (June 29, 2015)); see also *Fisher v. Univ. of Tex.*, 133 S. Ct. 2411, 2419 (2013) (“[A]ny official action that treats a person differently on account of his race or ethnic origin is inherently suspect.” (internal citation omitted)).

220. See *supra* Part I.A.

221. See David J. Hand, *Deconstructing Statistical Questions*, 157 J. ROYAL STAT. SOC’Y. SERIES A (STAT. SOC’Y) 317, 318–20 (1994).

222. Gandy, *supra* note 31, at 39.

innovation. Any argument for the superiority of one target variable over the other will simply make appeals to competing and incommensurate values.

For these same reasons, however, defining the target variable also offers an opportunity for creative thinking about the potentially infinite number of ways of making sound hiring decisions. Data miners can experiment with multiple definitions that each seem to serve the same goal, even if these fall short of what they themselves consider ideal. In principle, employers should rely on proxies that are maximally proximate to the actual skills demanded of the job. While there should be a tight nexus between the sought-after features and these skills, this may not be possible for practical and economic reasons. This leaves data miners in a position to dream up many different nonideal ways to make hiring decisions that may have a greater or less adverse impact on protected classes.

The Second Circuit considered such an approach in *Hayden v. County of Nassau*.<sup>223</sup> In *Hayden*, the county's goal was to find a police entrance exam that was "valid, yet minimized the adverse impact on minority applicants."<sup>224</sup> The county thus administered an exam with twenty-five parts that could be scored independently. By design, a statistically valid result could be achieved by one of several configurations that counted only a portion of the test sections, without requiring all of them.<sup>225</sup> The county ended up using nine of the sections as a compromise, after rejecting one configuration that was more advantageous to minority applicants but less statistically sound.<sup>226</sup> This is a clear example of defining a problem in such a way that it becomes possible to reduce the disparate impact without compromising the accuracy of the assessment mechanism.

## 2. Training Data

### a. Labeling Examples

Any solution to the problems presented by labeling must be a compromise between a rule that forbids employers from relying on past discrimination and one that allows them to base hiring decisions on historical examples of good employees. In theory, a rule that forbids employers from modeling decisions based on historical examples tainted by prejudice would address the problem of improper labeling. But if the only examples an employer has to draw on are those of past employees who had been subject to discrimination, all learned rules will recapitulate this discrimination.

Title VII has always had to balance its mandate to eliminate discrimination in the workplace with employers' legitimate discretion. For

---

223. 180 F.3d 42, 47 (2d Cir. 1999).

224. *Id.*

225. *Id.*

226. *Id.*

example, one of the most common selection procedures that explicitly reproduced past discrimination was seniority.<sup>227</sup> Seniority was, and is still often, a legitimate metric for promotion and is especially important in collective bargaining. After the passage of Title VII, however, seniority was also often used to keep black people from advancing to better jobs because they had not been hired until Title VII forced employers to hire them.<sup>228</sup> Despite this obvious problem with seniority, Title VII contains an explicit carve-out for “bona fide seniority or merit system[s].”<sup>229</sup> As a result, the Supreme Court has held that “absent a discriminatory purpose, the operation of a seniority system cannot be an unlawful employment practice even if the system has some discriminatory consequences.”<sup>230</sup> Given the inherent tension between ensuring that past discrimination is not reproduced in future decisions and permitting employers legitimate discretion, it should be unsurprising that, when translated to data mining, the problem is not amenable to a clear solution.

In fact, this difficulty is even more central to data mining. Data miners who attempt to remove the influence of prejudice on prior decisions by recoding or relabeling examples may find that they cannot easily resolve what the nonprejudicial determination would have been. As Calders and Žliobaitė point out, “[T]he notion of what is the correct label is fuzzy.”<sup>231</sup> Employers are unlikely to have perfectly objective and exhaustive standards for hiring; indeed, part of the hiring process is purposefully subjective. At the same time, employers are unlikely to have discriminated so completely in the past that the only explanation for rejecting an applicant was membership in protected classes. This leaves data miners tasked with correcting for prior prejudice with the impossible challenge of determining what the correct subjective employment decision would have been absent prejudice. Undoing the imprint of prejudice on the data may demand a complete re-rendering of the biased decisions rather than simply adjusting those decisions according to some fixed statistical measure.

### *b. Data Collection*

Although there are some cases with obviously skewed datasets that are relatively easy to identify and correct, often the source and degree of the bias will not be immediately apparent.<sup>232</sup> Street Bump suffered from a visually

---

227. Selmi, *supra* note 178, at 715.

228. See *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 450 (1975) (Burger, J., concurring) (“The basis of Albemarle’s liability was that its seniority system perpetuated the effects of past discrimination . . .”).

229. 42 U.S.C. § 2000e-2(h) (2012).

230. *Trans World Airlines, Inc. v. Hardison*, 432 U.S. 63, 82 (1977).

231. Calders & Žliobaitė, *supra* note 64, at 48.

232. For example, establishing whether and to what extent crime statistics misrepresent the relative proportion of offenses committed by different social groups is not an easy task. Especially challenging are those crimes that are more likely to go under- or unreported if not directly observed by

evident bias when the data was plotted on a map. Boston's Office of New Urban Mechanics was therefore able to partner with "a range of academics to take into account issues of equitable access and digital divides."<sup>233</sup> In many cases, however, an analyst can only determine the extent of—and correct for—unintentional discrimination that results from reporting, sampling, and selection biases if the analyst has access to information that somehow reveals misrepresentations of protected classes in the dataset. Often, there may be no practical alternative method for collecting information that would even reveal the existence of a bias.

Any attempt to correct for collection bias immediately confronts the problem of whether or not the employer recognizes the specific type of bias that is producing disparate results. Then, in order to correct for it, an employer must have access to the underlying data and often an ability to collect more. Where more data is clearly not accessible, data miners can proactively compensate for some of the bias by oversampling underrepresented communities.<sup>234</sup>

If the employer fails to be proactive or tries and fails to detect the bias that causes the disparate impact, liability is an open question. As discussed in Part II.B, liability partly depends on how liberally a court interprets the requirement that an employer "refuses" to use an alternative scheme.<sup>235</sup> Even a liberal interpretation, though, would require evidence of the particular type of discrimination at issue, coupled with evidence that such an alternative scheme exists. Thus, finding liability seems unlikely. Worse, where such showing is possible, there may be no easy or obvious way to remedy the situation.

To address collection bias directly, an employer or an auditor must have access to the underlying data and the ability to adjust the model. Congress could require this directly of any employer using data mining techniques. Some employers are investing in their own data now and could potentially meet such requirements.<sup>236</sup> But employers also seem happy to rely on models developed and administered by third parties, who may have a far greater set of examples and far richer data than any individual company.<sup>237</sup> Furthermore, due to economies of scale that are especially important in data analysis, one can imagine that third parties specializing in work-force science will be able to offer employers this service much less expensively than they could manage it

---

the police. See BERNARD E. HARCOURT, *AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE* (2007).

233. Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), <https://hbr.org/2013/04/the-hidden-biases-in-big-data> [https://perma.cc/9A7V-3UVD]. Such techniques would also address the concerns raised in Lerman, *supra* note 50.

234. Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE & INFO. SYS. 1, 3 (2011).

235. See *supra* Part II.B.

236. Peck, *supra* note 181.

237. See Richtel, *supra* note 69.



themselves. If Congress attempted to demand that employers have access to the data, it would face strong resistance from the ever-growing data analysis industry, whose business depends on the proprietary nature of the amassed information. More likely, Congress could require audits by a third party like the EEOC or a private auditor, in order to protect trade secrets, but this still seems a tall task. Ultimately, because proactive oversampling and retroactive data correction are at least possible, collection bias has the most promising prospects for a workable remedy of any of the identified data mining mechanisms.

### 3. *Feature Selection*

Even in the absence of prejudice or bias, determining the proper degree of precision in the distinctions drawn through data mining can be extremely difficult. Under formal disparate treatment, this is straightforward: any decision that expressly classifies by membership in a protected class is one that draws distinctions on illegitimate grounds. It is far less clear, however, what constitutes legitimate statistical discrimination when individuation does not rely on proscribed criteria. In these cases, the perceived legitimacy seems to depend on a number of factors: (1) whether the errors seem avoidable because (2) gaining access to additional or more granular data would be trivial or (3) would not involve costs that (4) outweigh the benefits. This seems to suggest that the task of evaluating the legitimacy of feature selection can be reduced to a rather straightforward cost-benefit analysis. Companies would have an obligation to pursue ever more—and more granular—data until the costs of gathering that data exceed the benefits conferred by the marginal improvements in accuracy.

Unfortunately, as is often the case with cost-benefit analyses, this approach fails to consider how different actors will perceive the value of the supposed benefits as well as the costs associated with errors. The obvious version of this criticism is that “actuarially saddled” victims of inaccurate determinations may find cold comfort in the fact that certain decisions are rendered more reliably overall when decision makers employ data mining.<sup>238</sup> A more sophisticated version of this criticism focuses on the way such errors assign costs and benefits to different actors at systematically different rates. A model with any error rate that continues to turn a profit may be acceptable to decision makers at a company, no matter the costs or inconvenience to specific customers.<sup>239</sup> Even when companies are subject to market pressures that would

---

238. SCHAUER, *supra* note 67, at 5. As Schauer explains, perfectly particularized decisions are, of course, a logical impossibility. Accepting this inherent limitation introduces a different sort of procedural concern: occasional errors might be tolerable if they are easy to detect and rectify, which is why, among other things, the perceived legitimacy of decisions often also depends on due process. *See id.* at 172; *see also* Citron, *supra* note 11.

239. Gandy, *supra* note 31, at 36.

force them to compete by lowering these error rates, the companies may find that there is simply no reason to invest in efforts that do so if the errors happen to fall disproportionately on especially unprofitable groups of consumers. Furthermore, assessing data mining as a matter of balancing costs and benefits leaves no room to consider morally salient disparities in the degree to which the costs are borne by different social groups. This raises the prospect that there might be systematic differences in the rates at which members of protected classes are subject to erroneous determinations.<sup>240</sup> Condemning these groups to bear the disproportionate burden of erroneous determinations would strike many as highly objectionable, despite greater accuracy in decision making for the majority group.<sup>241</sup> Indeed, simply accepting these cost differences as a given would subject those already in less favorable circumstances to less accurate determinations.

Even if companies assume the responsibility for ensuring that members of protected classes do not fall victim to erroneous determinations at systematically higher rates, they could find that increasing the resolution and range of their analyses still fails to capture the causal relationships that account for different outcomes because those relationships are not easily represented in data.<sup>242</sup> In such cases, rather than reducing the error rate for those in protected classes, data miners could structure their analyses to minimize the difference in error rates between groups. This solution may involve some unattractive tradeoffs, however. In reducing the disparate impact of errors, it may increase the overall amount of errors. In other words, generating a model that is equally unfair to protected and unprotected classes might increase the overall amount of unfairness.

#### 4. *Proxies*

Computer scientists have been unsure how to deal with redundant encodings in datasets. Simply withholding these variables from the data mining exercise often removes criteria that hold demonstrable and justifiable relevance to the decision at hand. As Calders and Žliobaitė note, “[I]t is problematic [to remove a correlated attribute] if the attribute to be removed also carries some objective information about the label [quality of interest].”<sup>243</sup> Part of the problem seems to be that there is no obvious way to determine *how* correlated a relevant attribute must be with class membership to be worrisome. Nor is there a self-evident way to determine when an attribute is sufficiently relevant to justify its consideration, despite its high correlation with class membership. As

---

240. Moritz Hardt, *How Big Data Is Unfair*, MEDIUM (Sept. 26, 2014), <https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de> [<https://perma.cc/YN44-M4DQ>].

241. See, e.g., Gandy, *supra* note 31, at 39.

242. See *supra* note 64 and accompanying text.

243. Calders & Žliobaitė, *supra* note 64, at 54.

Professors Devin Pope and Justin Sydnor explain, “[V]ariables are likely neither solely predictive nor purely proxies for omitted characteristics.”<sup>244</sup>

But there is a bigger problem here: attempting to ensure fairly rendered decisions by excising highly correlated criteria only makes sense if the disparate impact happens to be an *avoidable* artifact of a particular way of rendering decisions. And yet, even when denied access to these highly correlated criteria, data mining may suggest alternative methods for rendering decisions that still result in the same disparate impact. Focusing on isolated data points may be a mistake because class membership can be encoded in more than one specific and highly correlated criterion. Indeed, it is very likely that class membership is reflected across a number of interrelated data points.<sup>245</sup> But such outcomes might instead demonstrate something more unsettling: that *other* relevant criteria, whatever they are, happen to be possessed at different rates by members of protected classes. This explains why, for instance, champions of predictive policing have responded to critics by arguing that “[i]f you wanted to remove everything correlated with race, you couldn’t use anything. That’s the reality of life in America.”<sup>246</sup> Making accurate determinations means considering factors that are somehow correlated with proscribed features.

Computer scientists have even shown that “[r]emoving all such correlated attributes before training does remove discrimination, but with a high cost in classifier accuracy.”<sup>247</sup> This reveals a rather uncomfortable truth: the current distribution of relevant attributes—attributes that can and should be taken into consideration in apportioning opportunities fairly—is demonstrably correlated with sensitive attributes because the sensitive attributes have meaningfully conditioned what relevant attributes individuals happen to possess.<sup>248</sup> As such, attempts to ensure procedural fairness by excluding certain criteria from consideration may conflict with the imperative to ensure accurate determinations. The only way to ensure that decisions do not systematically disadvantage members of protected classes is to reduce the overall accuracy of

---

244. Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J. 206, 206 (2011).

245. *Supra* discussion accompanying note 101.

246. Labi, *supra* note 5 (quoting Ellen Kurtz, Director of Research for Philadelphia’s Adult Probation and Parole Department).

247. Toon Calders & Sicco Verwer, Presentation at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: Three Naïve Bayes Approaches for Discrimination-Free Classification 9 (2010), [http://www.wis.win.tue.nl/~tcalders/dadm/lib/exe/fetch.php?media=ecmlpkdd\\_2010\\_discrimination.pdf](http://www.wis.win.tue.nl/~tcalders/dadm/lib/exe/fetch.php?media=ecmlpkdd_2010_discrimination.pdf) [<http://perma.cc/9V72-2NVM>].

248. In a sense, computer scientists have unwittingly furnished the kind of evidence that social scientists routinely seek: the particular contours of inequality. *See, e.g.*, SOCIAL INEQUALITY (Kathryn M. Neckerman ed., 2004).

all determinations. As Dwork et al. remark, these results “demonstrate a quanti[t]ative trade-off between fairness and utility.”<sup>249</sup>

In certain contexts, data miners will never be able to fully disentangle legitimate and proscribed criteria. For example, the workforce optimization consultancy, Evolv, discovered that “[d]istance between home and work . . . is strongly associated with employee engagement and retention.”<sup>250</sup> Despite the strength of this finding, Evolv “never factor[s] [it] into the score given each applicant . . . because different neighborhoods and towns can have different racial profiles, which means that scoring distance from work could violate equal-employment-opportunity standards.”<sup>251</sup> Scholars have taken these cases as a sign that the “major challenge is how to find out which part of information carried by a sensitive (or correlated) attribute is sensitive and which is objective.”<sup>252</sup> While researchers are well aware that this may not be easy to resolve, let alone formalize into a computable problem, there is a bigger challenge from a legal perspective: any such undertaking would necessarily wade into the highly charged debate over the degree to which the relatively less favorable position of protected classes warrants the protection of antidiscrimination law in the first instance.

The problems that render data mining discriminatory are very rarely amenable to obvious, complete, or welcome resolution. When it comes to setting a target variable and feature selection, policy cannot lay out a clear path to improvement; reducing the disparate impact will necessitate open-ended exploration without any way of knowing when analysts have exhausted the possibility for improvement. Likewise, policies that compel institutions to correct tainted datasets or biased samples will make impossible demands of analysts. In most cases, they will not be able to determine what the objective determination should have been or independently observe the makeup of the entire population. Dealing with both of these problems will ultimately fall to analysts’ considered judgment. Solutions that reduce the accuracy of decisions to minimize the disparate impact caused by coarse features and unintentional proxies will force analysts to make difficult and legally contestable trade-offs. General policies will struggle to offer the specific guidance necessary to determine the appropriate application of these imperfect solutions. And even when companies voluntarily adopt such strategies, these internal difficulties will likely allow a disparate impact to persist.

---

249. Dwork et al., *supra* note 81, at 215; *cf.* Wax, *supra* note 146, at 711 (noting intractable problems due to a “validity-diversity tradeoff” in employment metrics).

250. Peck, *supra* note 181.

251. *Id.* Other companies have not held back from considering this information for the very same purposes. See Joseph Walker, *Meet the New Boss: Big Data*, WALL ST. J. (Sept. 20, 2012), <http://www.wsj.com/news/articles/SB10000872396390443890304578006252019616768> [<https://perma.cc/6DHY-M429>].

252. Calders & Žliobaitė, *supra* note 64, at 56.

### B. External Difficulties

Assuming the internal difficulties can be resolved, there are further political and constitutional restraints on addressing Title VII's inadequacies with respect to data mining. Data mining discrimination will force a confrontation between the two divergent principles underlying antidiscrimination law: anticlassification and antisubordination.<sup>253</sup> Which of these two principles motivates discrimination law is a contentious debate, and making remedies available under antidiscrimination law will require a commitment to antisubordination principles that have thus far not been forthcoming from legislatures. This is not merely a political concern, as substantive remediation is becoming ever more suspect constitutionally as well.<sup>254</sup> While such remedies may be politically and legally impossible, the nature of data mining itself makes them practically necessary. Accordingly, these external difficulties may prevent antidiscrimination law from fully addressing data mining discrimination.

Two competing principles have always undergirded antidiscrimination law: anticlassification and antisubordination. Anticlassification is the narrower of the two, holding that the responsibility of the law is to eliminate the unfairness individuals in certain protected classes experience due to decision makers' choices.<sup>255</sup> Antisubordination theory, in contrast, holds that the goal of antidiscrimination law is, or at least should be, to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance.<sup>256</sup>

Different mitigation policies effectuate different rationales. Disparate treatment doctrine arose first, clearly aligning with the anticlassification principle by proscribing intentional discrimination, in the form of either explicit singling out of protected classes for harm or masked intentional discrimination. Since disparate impact developed, however, there has never been clarity as to which of the principles it is designed to effectuate.<sup>257</sup> On the one hand, disparate impact doctrine serves anticlassification by being an "evidentiary dragnet" used to "smoke out" well-hidden disparate treatment.<sup>258</sup> On the other hand, as an effects-based doctrine, there is good reason to believe it was intended to address substantive inequality.<sup>259</sup> In this sense, the "business

---

253. Helen Norton, *The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality*, 52 WM. & MARY L. REV. 197, 206–15 (2010); see also Bagenstos, *supra* note 122, at 40–42, 40–41 nn.214–15 (collecting sources); Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 157 (1976).

254. See Norton, *supra* note 253.

255. *Id.* at 209.

256. *Id.* at 206.

257. Primus, *supra* note 98, at 520–23.

258. *Id.*; Perry, *supra* note 138, at 526.

259. See *Griggs v. Duke Power Co.*, 401 U.S. 424, 429–30 (1971) ("The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to achieve equality of

necessity” defense is a necessary backstop that prevents members of traditionally disadvantaged groups from simply forcing their way in without the necessary skills or abilities.<sup>260</sup>

Thus, the mapping from anticlassification and antisubordination to disparate treatment and disparate impact was never clean. Early critics of civil rights laws actually complained that proscribing consideration of protected class was a subsidy to black people.<sup>261</sup> This argument quickly gave way in the face of the rising importance of the anticlassification norm.<sup>262</sup> Over the years, the anticlassification principle has come to dominate the landscape so thoroughly that a portion of the populace thinks (as do a few Justices on the Supreme Court) that it is the only valid rationale for antidiscrimination law.<sup>263</sup>

The move away from antisubordination began only five years after disparate impact was established in *Griggs*. In *Washington v. Davis*, the Court held that disparate impact could not apply to constitutional claims because equal protection only prohibited intentional discrimination.<sup>264</sup> Since then, the various affirmative action cases have overwritten the distinction between benign and harmful categorizations of race in favor of a formalistic anticlassification principle, removed from its origins as a tool to help members of historically disadvantaged groups.<sup>265</sup> White men can now bring disparate treatment claims.<sup>266</sup> If antidiscrimination law is no longer thought to serve the purpose of improving the relative conditions of traditionally disadvantaged groups, antisubordination is not part of the equation.

While the Court has clearly established that antisubordination is not part of constitutional equal protection doctrine, that it does not mean that antisubordination cannot animate statutory antidiscrimination law. Antisubordination and anticlassification came into sharp conflict in *Ricci v. DeStefano*, a 2009 case in which the City of New Haven refused to certify a promotion exam given to its firefighters on the grounds that it would have produced a disparate impact based on its results.<sup>267</sup> The Supreme Court held that the refusal to certify the test, a facially race-neutral attempt to correct for perceived disparate impact, was in fact a race-conscious remedy that constituted disparate treatment of the majority-white firefighters who would

---

employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to ‘freeze’ the status quo of prior discriminatory employment practices.”).

260. See *Tex. Dep’t of Hous. & Cmty Aff. v. Inclusive Cmty. Project, Inc.*, No. 13-1371, slip op. at 8 (Sup. Ct. 2015) (quoting *Griggs*, 401 U.S. at 431).

261. *Primus*, *supra* note 98, at 525–26.

262. *Id.*

263. See *Bagenstos*, *supra* note 122, at 41.

264. 426 U.S. 229, 246–48 (1976).

265. *Rubinfeld*, *supra* note 99, at 428, 433–36.

266. *Ricci v. DeStefano*, 557 U.S. 557 (2009).

267. *Id.*

have been promoted based on the exam's results.<sup>268</sup> The Court held that disparate treatment cannot be a remedy for disparate impact without a "strong basis in evidence" that the results would lead to actual disparate treatment liability.<sup>269</sup>

*Ricci* was the first indication at the Supreme Court that disparate impact doctrine could be in conflict with disparate treatment.<sup>270</sup> The Court had previously ruled in essence that the antisubordination principle could not motivate a constitutional decision,<sup>271</sup> but it had not suggested that law effectuating that principle could itself be discriminatory against the dominant groups. That has now changed.<sup>272</sup>

The decision has two main consequences for data mining. First, where the internal difficulties in resolving discrimination in data mining described above can be overcome, legislation that requires or enables such resolution may run afoul of *Ricci*. Suppose, for example, Congress amended Title VII to require that employers make their training data and models auditable. In order to correct for detected biases in the training data that result in a model with a disparate impact, the employer would first have to consider membership in the protected class. The remedy is inherently race-conscious. The *Ricci* Court did hold that an employer may tweak a test during the "test-design stage," however.<sup>273</sup> So, as a matter of timing, data mining might not formally run into

---

268. *Id.*

269. *Id.* at 563.

270. Primus, *supra* note 92, at 1344; Lawrence Rosenthal, *Saving Disparate Impact*, 34 CARDOZO L. REV. 2157, 2162–63 (2013); Norton, *supra* note 253, at 229.

271. *See* *Washington v. Davis*, 426 U.S. 229, 239 (holding that discriminatory purpose is necessary to finding a violation of equal protection).

272. Primus, *supra* note 92, at 1343. While the decision was formally about Title VII only, and thus amenable to statutory resolution, the reasoning applied equally well to a future equal protection claim, endangering the future of disparate impact. *Id.* at 1385–87; Bradley A. Areheart, *The Anticlassification Turn in Employment Discrimination Law*, 63 ALA. L. REV. 955, 994 (2012); Norton, *supra* note 253, at 229–30. Justice Scalia stated as much in his concurrence. *Ricci*, 557 U.S. at 594 (Scalia, J., concurring) ("[The Court's] resolution of this dispute merely postpones the evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution's guarantee of equal protection?"). But the Supreme Court seemed to pull back from the brink last term, approving of the use of disparate impact in a new setting—the Fair Housing Act—and engaging deeply with the constitutional issues that *Ricci* raised, settling them for now. Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities*, 101 CORNELL L. REV., at \*11–12 (forthcoming 2016), [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2642631](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2642631) [https://perma.cc/WD43-XW2G]; Richard Primus, *Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact After Ricci and Inclusive Communities*, in *TITLE VII OF THE CIVIL RIGHTS ACT AFTER 50 YEARS: PROCEEDINGS OF THE NEW YORK UNIVERSITY 67TH ANNUAL CONFERENCE ON LABOR* 295 (2015).

273. *Ricci*, 557 U.S. at 585 (majority opinion) ("Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race. And when, during the test-design stage, an employer invites comments to ensure the test is fair, that process can provide a common ground for open discussions toward that end.").

*Ricci* if the bias resulting in a disparate impact is corrected before applied to individual candidates. After an employer begins to use the model to make hiring decisions, only a “strong basis in evidence” that the employer will be successfully sued for disparate impact will permit corrective action.<sup>274</sup> Of course, unless every single model used by an employer is subject to a prescreening audit (an idea that seems so resource intensive that it is effectively impossible), the disparate impact will be discovered only when the employer faces complaints. Additionally, while *Ricci*’s holding was limited in scope, the “strong basis in evidence” standard did not seem to be dictated by the logic of the opinion, which illustrated a more general conflict between disparate treatment and disparate impact.<sup>275</sup>

Second, where the internal difficulties *cannot* be overcome, there is likely no way to correct for the discriminatory outcomes aside from results-focused balancing, and requiring this will pose constitutional problems. For those who adhere to the anticlassification principle alone, such an impasse may be perfectly acceptable. They might say that as long as employers are not intentionally discriminating based on explicitly proscribed criteria, the chips should fall where they may. To those who believe some measure of substantive equality is important over and above procedural equality, this result will be deeply unsatisfying.

An answer to the impasse created by situations that would require results-focused rebalancing is to reexamine the purpose of antidiscrimination law. The major justification for reliance on formal disparate treatment is that prejudice is simply irrational and thus unfair. But if an employer knows that his model has a disparate impact, but it is also his most predictive, the argument that the discrimination is irrational loses any force. Thus, data mining may require us to reevaluate why and whether we care about not discriminating.

Consider another example involving tenure predictions, one in which an employer ranks potential employees with the goal of hiring only those applicants that the company expects to retain for longer periods of time. In optimizing its selection of applicants in this manner, the employer may unknowingly discriminate against women if the historical data demonstrates that they leave their positions after fewer years than their male counterparts. If gender accounts for a sufficiently significant difference in employee tenure, data mining will generate a model that simply discriminates on the basis of gender or those criteria that happen to be proxies for gender. Although selecting applicants with an eye to retention might seem both rational and reasonable, granting significance to predicted tenure would subject women to systematic disadvantage if gender accounts for a good deal of the difference in tenure. If that is the case, any data mining exercise that attempts to predict

---

274. *Id.* at 585.

275. *See generally id.*



tenure will invariably rediscover this relationship. One solution could be for Congress to amend Title VII to reinvigorate strict business necessity.<sup>276</sup> This would allow a court to accept that relying on tenure is rational but not strictly “necessary” and that perhaps other factors could make up for the lack of predicted tenure.

But this solution and all others must rely on the antidisubordination principle. Consider this question: should the law permit a company to hire no women at all—or none that it correctly predicts will depart following the birth of a child—because it is the most rational choice according to their model?<sup>277</sup> The answer seems obviously to be no. But why not? What forms the basis for law’s objection to rational decisions, based on seemingly legitimate criteria, that place members of protected classes at systematic disadvantage? The Supreme Court has observed that, “Title VII requires employers to treat their employees as *individuals*, not ‘as simply components of a racial, religious, sexual, or national class.’”<sup>278</sup> On the strength of that statement, the Court held that employers could not force women to pay more into an annuity because they, as women, were likely to live longer.<sup>279</sup> But it is not clear that this reasoning translates directly to data mining. Here, the model takes a great deal of data about an individual, and while it does make a determination based on statistics, it will make a different one if analyzing two different women. So if the model said to hire *no* women, it would be illegal, but, according to the doctrine, perhaps only because every woman ends up with the same result.

The only escape from this situation may be one in which the relevance of gender in the model is purposefully ignored and all factors correlated with gender are suppressed. The outcome would be a necessarily less accurate model. The justification for placing restrictions on employers, and limiting the effectiveness of their data mining, would have to depend on an entirely different set of arguments than those advanced to explain the wrongfulness of biased data collection, poorly labeled examples, or an impoverished set of features. Here, shielding members of protected classes from less favorable treatment is not justified by combatting prejudice or stereotyping. In other words, any prohibition in this case could not rest on a procedural commitment to ensuring ever more accurate determinations. Instead, the prohibition would have to rest on a substantive commitment to equal representation of women in the workplace. That is, it would have to rest on a principle of antidisubordination.

---

276. Remember that if there is disparate impact, but no liability, it is because the goal was deemed job-related or satisfied business necessity.

277. As a matter of case law, this question has essentially been answered. The Supreme Court has ruled that in the case of women being required to pay more into an annuity because they would likely live longer, pure market rationality is not a good enough answer. *Ariz. Governing Comm. v. Norris*, 463 U.S. 1073, 1083 (1983) (quoting *City of Los Angeles Dep’t of Water & Power v. Manhart*, 435 U.S. 702, 708 (1978)).

278. *Id.*

279. *Id.*

The dilemma is clear: the farther the doctrine gets from substantive remediation, the less utility it has in remedying these kinds of discriminatory effects.<sup>280</sup> But the more disparate impact is thought to embody the antistatutory principle—as opposed to the “evidentiary dragnet” in service of the antistatutory norm—the more it will invite future constitutional challenges.<sup>281</sup>

This also raises a point about disparate *treatment* and data mining. Within data mining, the effectiveness of prohibiting the use of certain information exists on a spectrum. On one end, the prohibition has little to no effect because either the information is redundantly encoded or the results do not vary along lines of protected class. On the other end, the prohibition reduces the accuracy of the models. That is, if protected class data were not prohibited, that information would alter the results, presumably by making members of protected classes worse (or, in some cases, better) off. Thus, as a natural consequence of data mining, a command to ignore certain data has either no effect<sup>282</sup> or the effect of altering the fortunes of those protected classes in substantive ways. Therefore, with respect to data mining, due to the zero-sum nature of a ranking system, even *disparate treatment* doctrine is a reallocative remedy similar to affirmative action.<sup>283</sup> Once again, this erodes the legitimate rationale for on the one hand supporting an antistatutory principle but on the other, holding fast against antistatutory in this context. The two principles tend to accomplish the same thing, but one is less effective at achieving substantive equality.

This reveals that the pressing challenge does not lie with ensuring procedural fairness through a more thorough stamping out of prejudice and bias but rather with developing ways of reasoning to adjudicate when and what amount of disparate impact is tolerable. Abandoning a belief in the efficacy of procedural solutions leaves policy makers in an awkward position because there is no definite or consensus answer to questions about the fairness of specific outcomes. These need to be worked out on the basis of different normative principles. At some point, society will be forced to acknowledge that this is really a discussion about what constitutes a tolerable level of disparate impact in employment. Under the current constitutional order and in the political climate, it is tough to even imagine having such a conversation. But, until that happens, data mining will be permitted to exacerbate existing inequalities in difficult-to-counter ways.

---

280. *Id.* at 537.

281. Primus, *supra* note 98, at 536–37.

282. *See supra* text accompanying note 101.

283. For an argument that this is true more generally, see Bagenstos, *supra* note 90, and Owen M. Fiss, *A Theory of Fair Employment Laws*, 38 U. CHI. L. REV. 235, 313 (1971) (arguing that a key to understanding antidiscrimination prohibitions in the employment realm is that the prohibitions “confer[] benefits on a racial class—blacks”).

## CONCLUSION

This Essay has identified two types of discriminatory outcomes from data mining: a family of outcomes where data mining goes “wrong” and outcomes where it goes too “right.” Data mining can go wrong in any number of ways. It can choose a target variable that correlates to protected class more than others would, reproduce the prejudice exhibited in the training examples, draw adverse lessons about protected classes from an unrepresentative sample, choose too small a feature set, or not dive deep enough into each feature. Each of these potential errors is marked by two facts: the errors may generate a manifest disparate impact, and they may be the result of entirely innocent choices made by data miners.

Where data mining goes “right,” data miners could not have been any more accurate given the starting point of the process. This very accuracy, exposing an uneven distribution of attributes that predict the target variable, gives such a result its disparate impact. If the data accurately models inequality, attempts to devise an alternative way of making the same prediction will only narrow the disparate impact if these efforts reduce the accuracy of the decision procedure. By now, it should be clear that Title VII, and very likely other similarly process-oriented civil rights laws, cannot effectively address this situation.

This means something different for the two families, and it should be slightly more surprising for the former. At a high level of abstraction, where a decision process goes “wrong” and this wrongness creates a disparate impact, Title VII and similar civil rights laws should be up to the task of solving the problem; that is ostensibly their entire purpose. But aside from a few more obvious cases involving manifest biases in the dataset, it is quite difficult to determine ahead of time what “correct” data mining looks like. A decision maker can rarely discover that the choice of a particular target variable is more discriminatory than other choices until after the fact, at which point it may be difficult and costly to change course. While data miners might have some intuitions about the influence that prejudice or bias played in the prior decisions that will serve as training data, data miners may not have any systematic way of measuring and correcting for that influence. And even though ensuring reliable samples before training a model is a possibility, the data may never be perfect. It may be impossible to determine, *ex ante*, how much the bias contributes to the disparate impact, it may not be obvious how to collect additional data that makes the sample more representative, and it may be prohibitively expensive to do so. Companies will rarely be able to resolve these problems completely; their models will almost always suffer from some deficiency that results in a disparate impact. A standard that holds companies liable for any amount of theoretically avoidable disparate impact is likely to ensnare all companies. Thus, even at this level of abstraction, it becomes clear that holding the decision makers responsible for these disparate impacts is at

least partly troubling from a due process perspective. Such concerns may counsel against using data mining altogether. This would be a perverse outcome, given how much even imperfect data mining can do to help reduce the very high rates of discrimination in employment decisions.

If liability for getting things “wrong” is difficult to imagine, how does liability for getting things “right” make any more sense? That proxy discrimination largely rediscovers preexisting inequalities suggests that perhaps Title VII is not the appropriate remedial vehicle. If what is at stake are the results of decades of historical discrimination and wealth concentration that have created profound inequality in society, is that not too big a problem to remedy through individual lawsuits, assuming affirmative action and similar policies are off the table? Thus, perfect data mining forces the question: if employers can say with certainty that, given the status quo,<sup>284</sup> candidates from protected classes are on average less ready for certain jobs than more privileged candidates, should employers specifically be penalized for hiring fewer candidates from protected classes?

Doctrinally, the answer is yes, to some extent. Professor Christine Jolls has written that disparate impact doctrine is akin to accommodation in disability law—that is, both accommodations and disparate impact specifically require employers to depart from pure market rationality and incur costs associated with employing members of protected classes.<sup>285</sup> Similarly, the Title VII annuity cases<sup>286</sup> and Title VII’s ban on following racist third-party preferences<sup>287</sup> each require a departure from market rationality. Thus, Title VII makes that decision to a degree. But to what degree? How much cost must an employer bear?

Title VII does not require an employer to use the least discriminatory means of running a business.<sup>288</sup> Likewise, Title VII does not aim to remedy historical discrimination and current inequality by imposing all the costs of restitution and redistribution on individual employers.<sup>289</sup> It is more appropriately understood as a standard of *defensible* disparate impact. One route, then, to addressing the problems is to make the inquiry more searching and put the burden on the employer to avoid at least the easy cases. In a system that is as unpredictable as data mining can be, perhaps the proper way of

---

284. We cannot stress enough the import of these caveats. Certainty is a strong and unlikely precondition, and the status quo should not be taken as a given, as we explain below.

285. See generally Jolls, *supra* note 90.

286. See *Ariz. Governing Comm. v. Norris*, 463 U.S. 1073, 1083 (1983); *City of Los Angeles Dep’t of Water & Power v. Manhart*, 435 U.S. 702, 708 (1978).

287. See 29 C.F.R. § 1604.2(a)(1)(iii) (2015) (stating the EEOC’s position that “the preferences of coworkers, the employer, clients or customers” cannot be used to justify disparate treatment).

288. See, e.g., *El v. Se. Pa. Transp. Auth.*, 479 F.3d 232, 242 (3d Cir. 2007).

289. See Steven L. Willborn, *The Disparate Impact Model of Discrimination: Theory and Limits*, 34 AM. U. L. REV. 799, 809–10 (1985).

thinking about the solution is a duty of care, a theory of negligent discrimination.<sup>290</sup>

But if Title VII alone cannot solve these problems, where should society look for answers? Well, the first answer is to question the status quo. Data mining takes the existing state of the world as a given and ranks candidates according to their predicted attributes in *that* world. Data mining, by its very nature, treats the target variable as the only item that employers are in a position to alter; everything else that happens to correlate with different values for the target variable is assumed stable. But there are many reasons to question these background conditions. Sorting and selecting individuals according to their apparent qualities hides the fact that the predicted effect of possessing these qualities with respect to a specific outcome is also a function of the conditions under which these decisions are made. Recall the tenure example from Part III.B. In approaching appropriate hiring practices as a matter of selecting the “right” candidates at the outset, an employer will fail to recognize potential changes that he could make to workplace conditions. A more family-friendly workplace, greater on-the-job training, or a workplace culture more welcoming to historically underrepresented groups could affect the course of employees’ tenure and their long-term success in ways that undermine the seemingly prophetic nature of data mining’s predictions.

These are all traditional goals for reducing discrimination within the workplace, and they continue to matter even in the face of the eventual widespread adoption of data mining. But data can play a role here, too. For example, comparing the performance of equally qualified candidates across different workplaces can help isolate the formal policies and institutional dynamics that are more or less likely to help workers flourish. Research of this sort could also reveal areas for potential reform.<sup>291</sup>

Education is also important. Employers may take some steps to rectify the problem on their own if they better understand the cause of the disparity. Right now, many of the problems described in Part I are relatively unknown. But the more employers and data miners understand these pitfalls, the more they can strive to create better models on their own. Many employers switch to data-driven practices for the express purpose of eradicating bias;<sup>292</sup> if employers discover that they are introducing new forms of bias, they can correct course.

Even employers seeking only to increase efficiency or profit may find that their incentives align with the goals of nondiscrimination. Faulty data and data

---

290. See generally David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899 (1993).

291. Solon Barocas, *Putting Data to Work*, DATA AND DISCRIMINATION: COLLECTED ESSAYS 58, 60 (Seeta Peña Gangadharan, Virginia Eubanks & Solon Barocas eds., 2014).

292. Claire Cain Miller, *Can an Algorithm Hire Better than a Human?*, N.Y. TIMES (June 25, 2015), <http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html> [https://perma.cc/UR37-83D4].

mining will lead employers to overlook or otherwise discount people who are actually “good” employees. Where the cost of addressing these problems is at least compensated for by a business benefit of equal or greater value, employers may have natural incentives to do so.

Finally, employers could also make more effective use of the tools that computer scientists have begun to develop.<sup>293</sup> Advances in these areas will depend, crucially, on greater and more effective collaboration between employers, computer scientists, lawyers, advocates, regulators, and policy makers.<sup>294</sup>

This Essay is a call for caution in the use of data mining, not its abandonment. While far from a panacea, data mining can and should be part of a panoply of strategies for combatting discrimination in the workplace and for promoting fair treatment and equality. Ideally, institutions can find ways to use data mining to generate new knowledge and improve decision making that serves the interests of both decision makers and protected classes. But where data mining is adopted and applied without care, it poses serious risks of reproducing many of the same troubling dynamics that have allowed discrimination to persist in society, even in the absence of conscious prejudice.

---

293. See list *supra* note 217.

294. Joshua A. Kroll, et al., *Accountable Algorithms*, 165 U. PA. L. REV. \_\_ (forthcoming 2017).

# Fair prediction with disparate impact: A study of bias in recidivism prediction instruments

Alexandra Chouldechova \*

Last revised: February 2017

## Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

**Keywords:** disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

## 1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing [1, 2, 3]. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If the use of RPI's is to become commonplace, it is especially important to ensure that the instruments are free from discriminatory biases that could result in unethical practices and inequitable outcomes for different groups.

In a recent widely popularized investigation conducted by a team at ProPublica, Angwin et al. [4] studied an RPI called COMPAS<sup>a</sup>, concluding that it is biased against black defendants. The

---

\*Heinz College, Carnegie Mellon University

<sup>a</sup>COMPAS [5] is a risk assessment instrument developed by Northpointe Inc.. Of the 22 scales that COMPAS provides, the Recidivism risk and Violent Recidivism risk scales are the most widely used. The empirical results in this paper are based on decile scores coming from the COMPAS Recidivism risk scale.

authors found that the likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of white defendants. Similarly, the likelihood of a recidivating black defendant being assessed as low risk is nearly half that of white defendants. In technical terms, these findings indicate that the COMPAS instrument has considerably higher false positive rates and lower false negative rates for black defendants than for white defendants.

ProPublica’s analysis has met with much criticism from both the academic community and from the Northpointe corporation. Much of the criticism has focussed on the particular choice of fairness criteria selected for the investigation. Flores et al. [6] argue that the correct approach for assessing RPI bias is instead to check for *calibration*, a fairness criterion that they show COMPAS satisfies. Northpointe in their response[7] argue for a still different approach that checks for a fairness criterion termed *predictive parity*, which they demonstrate COMPAS also satisfies. We provide precise definitions and a more in-depth discussion of these and other fairness criteria in Section 2.1.

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al. [4] are a direct consequence of applying an RPI that that satisfies predictive parity to a population in which recidivism prevalence<sup>a</sup> differs across groups. Our main contribution is twofold. (1) First, we make precise the connection between the predictive parity criterion and error rates in classification. (2) Next, we demonstrate how using an RPI that has different false positive and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties. Throughout our discussion we use the term *disparate impact* to refer to settings where a penalty policy has unintended disproportionate adverse impact on a particular group.

It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. A risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used. In this paper we consider hypothetical use cases in which we are able to directly connect particular fairness properties of an RPI to a measure of disparate impact. We present both theoretical and empirical results to illustrate how disparate impact can arise.

## 1.1 Outline of paper

We begin in Section 2 by providing some background on several of the different fairness criteria that have appeared in recent literature. We then proceed to demonstrate that an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups. In Section 3 we analyse a simple risk assessment-based sentencing policy and show how differences in false positive and false negative rates can result in disparate impact under this policy. In Section 3.3 we back up our theoretical analysis by presenting some empirical results based on the data made available by the ProPublica investigators. We conclude with a discussion of the issues that biased data presents for the arguments put forth in this paper.

---

<sup>a</sup>*Prevalence*, also termed the *base rate*, is the proportion of individuals who recidivate in a given population.



## 1.2 Data description and setup

The empirical results in this paper are based on the Broward County data made publicly available by ProPublica[8]. This data set contains COMPAS recidivism risk decile scores, 2-year recidivism outcomes, and a number of demographic and crime-related variables on individuals who were scored in 2013 and 2014. We restrict our attention to the subset of defendants whose race is recorded as African-American ( $b$ ) or Caucasian ( $w$ ).<sup>a</sup> After applying the same data pre-processing and filtering as reported in the ProPublica analysis, we are left with a data set on  $n = 6150$  individuals, of whom  $n_b = 3696$  are African-American and  $n_c = 2454$  are Caucasian.

## 2 Assessing fairness

### 2.1 Background

We begin by with some notation. Let  $S = S(x)$  denote the risk score based on covariates  $X = x \in \mathbb{R}^p$ , with higher values of  $S$  corresponding to higher levels of assessed risk. We will interchangeably refer to  $S$  as a *score* or an *instrument*. For simplicity, our discussion of fairness criteria will focus on a setting where there exist just two groups. We let  $R \in \{b, w\}$  denote the group to which an individual belongs, and do not preclude  $R$  from being one of the elements of  $X$ . We denote the outcome indicator by  $Y \in \{0, 1\}$ , with  $Y = 1$  indicating that the given individual goes on to recidivate. Lastly, we introduce the quantity  $s_{\text{HR}}$ , which denotes the high-risk score threshold. Defendants whose score  $S$  exceeds  $s_{\text{HR}}$  will be referred to as *high-risk*, while the remaining defendants will be referred to as *low-risk*.

With this notation in hand, we now proceed to define and discuss several fairness criteria that commonly appear in the literature, beginning with those mentioned in the introduction. We indicate cases where a given criterion is known to us to also commonly appear under some other name. All of the criteria presented below can also be assessed *conditionally* by further conditioning on some covariates in  $X$ . We discuss this point in greater detail in Section 3.1.

**Definition 1** (Calibration). A score  $S = S(x)$  is said to be *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals’ group membership. That is, if for all values of  $s$ ,

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w). \tag{2.1}$$

Within the educational and psychological testing and assessment literature, the notion of *calibration* features among the widely accepted and adopted standards for empirical fairness assessment. In this literature, an instrument that is *well-calibrated* is referred to as being *free from predictive bias*. This criterion has recently been applied to the PCRA<sup>b</sup> instrument, with initial findings suggesting that calibration is satisfied with respect race[10, 11], but not with respect to

---

<sup>a</sup>There are 6 racial groups represented in the data. 85% of individuals are either African-American or Caucasian.

<sup>b</sup>The Post Conviction Risk Assessment (PCRA) tool was developed by the Administrative Office of the United States Courts for the purpose of improving “the effectiveness and efficiency of post-conviction supervision” [9]

gender[12]. In their response to the ProPublica investigation, Flores et al. [6] verify that COMPAS is well-calibrated using logistic regression modeling.

**Definition 2** (Predictive parity). A score  $S = S(x)$  satisfies *predictive parity* at a threshold  $s_{\text{HR}}$  if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if,

$$\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = w). \quad (2.2)$$

Predictive parity at a given threshold  $s_{\text{HR}}$  amounts to requiring that the *positive predictive value* (PPV) of the classifier  $\hat{Y} = \mathbb{1}_{S > s_{\text{HR}}}$  be the same across groups. While predictive parity and calibration look like very similar criteria, well-calibrated scores can fail to satisfy predictive parity at a given threshold. This is because the relationship between (2.2) and (2.1) depends on the conditional distribution of  $S \mid R = r$ , which can differ across groups in ways that result in PPV imbalance. In the simple case where  $S$  itself is binary, a score that is well-calibrated will also satisfy predictive parity. Northpointe’s refutation[7] of the ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choices of interest.

**Definition 3** (Error rate balance). A score  $S = S(x)$  satisfies *error rate balance* at a threshold  $s_{\text{HR}}$  if the false positive and false negative error rates are equal across groups. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = w), \quad \text{and} \quad (2.3)$$

$$\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = w), \quad (2.4)$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates.

ProPublica’s analysis considered a threshold of  $s_{\text{HR}} = 4$ , which they showed leads to considerable imbalance in both false positive and false negative rates. While this choice of cutoff met with some criticism, we will see later in this section that error rate imbalance persists—indeed, must persist—for any choice of cutoff at which the score satisfies the predictive parity criterion. Error rate balance is also closely connected to the notions of *equalized odds* and *equal opportunity* as introduced in the recent work of Hardt et al. [13].

**Definition 4** (Statistical parity). A score  $S = S(x)$  satisfies *statistical parity* at a threshold  $s_{\text{HR}}$  if the proportion of individuals classified as high-risk is the same for each group. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid R = b) = \mathbb{P}(S > s_{\text{HR}} \mid R = w) \quad (2.5)$$

Statistical parity also goes by the name of *equal acceptance rates*[14] or *group fairness*[15], though it should be noted that these terms are in many cases not used synonymously. While our discussion focusses primarily on first three fairness criteria, statistical parity is widely used within the machine learning community and may be the criterion with which many readers are most familiar[16, 17]. Statistical parity is well-suited to contexts such as employment or admissions, where it may be desirable or required by law or regulation to employ or admit individuals in equal proportion across racial, gender, or geographical groups. It is, however, a difficult criterion to motivate in the recidivism prediction setting, and thus will not be further considered in this work.

## 2.2 Further related work

Though the study of discrimination in decision making and predictive modeling is rapidly evolving, it also has a long and rich multidisciplinary history. Romei and Ruggieri [18] provide an excellent overview of some of the work in this broad subject area. The recent work of Barocas and Selbst [19] offers a broad examination of algorithmic fairness framed within the context of anti-discrimination laws governing employment practices. Hannah-Moffat [20], Skeem [21], and Monahan and Skeem [22] examine legal and ethical issues relating specifically to the use of risk assessment instruments in sentencing, citing the potential for race and gender discrimination as a major concern.

In work concurrent with our own, several other researchers have also investigated the compatibility of different notions of fairness. Kleinberg et al. [23] show that calibration cannot be satisfied simultaneously with the fairness criteria of *balance for the negative class* and *balance for the positive class*. Translated into the present context, the latter criteria require that the average score assigned to non-recidivists (the negative class) should be the same for both groups, and that the same should hold among recidivists (the positive class). The work of Corbett-Davies et al. [24] closely parallels the results that we present in Section 2.3, reaching the same conclusion regarding the incompatibility of predictive parity and error rate balance in the setting of unequal prevalence.

## 2.3 Predictive parity, false positive rates, and false negative rates

In this section we present our first main result, which establishes that predictive parity is incompatible with error rate balance when prevalence differs across groups. To better motivate the discussion, we begin by presenting an empirical fairness assessment of the COMPAS RPI. Figure 1 shows plots of the observed recidivism rates and error rates corresponding to the fairness notions of calibration, predictive parity, and error rate balance. We see that the COMPAS RPI is (approximately) well-calibrated, and also satisfies predictive parity provided that the high-risk cutoff  $s_{HR}$  is 4 or greater. However, COMPAS fails on both false positive and false negative error rate balance across the range of high-risk cutoffs.

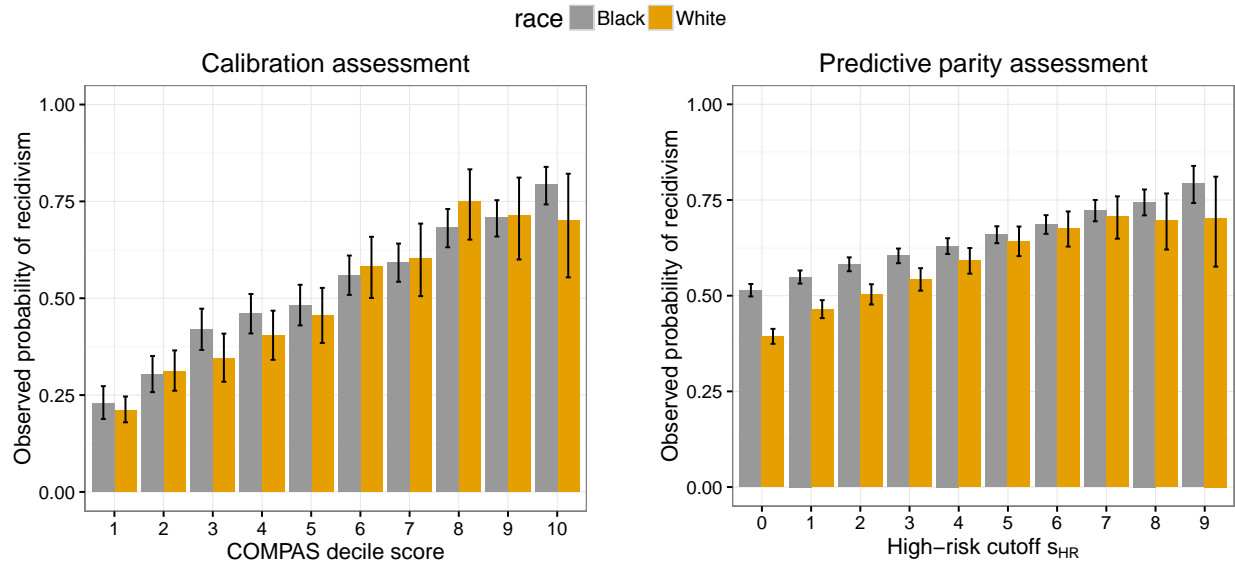
Angwin et al. [4] focussed on a high-risk cutoff of  $s_{HR} = 4$  for their analysis, which some critics have argued is too low, suggesting that  $s_{HR} = 7$  is more suitable. As can be seen from Figures 1c and 1d, significant error rate imbalance persists at this cut-off as well. Moreover, the error rates achieved at so high a cutoff are at odds with evidence suggesting that the use of RPI’s is of interest in settings where false negatives have a higher cost than false positives, with relative cost estimates ranging from 2.6 to upwards of 15. [25, 26]

As we now proceed to show, the error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context. When the recidivism prevalence—i.e., the base rate  $\mathbb{P}(Y = 1 \mid R = r)$ —differs across groups, any instrument that satisfies predictive parity at a given threshold  $s_{HR}$  *must* have imbalanced false positive or false negative errors rates at that threshold. To understand why predictive parity and error rate balance are mutually exclusive in the setting of unequal recidivism prevalence, it is instructive to think of how these quantities are all related.

Given a particular choice of  $s_{HR}$ , we can summarize an instrument’s performance in terms of a confusion matrix, as shown in Table 1 below.

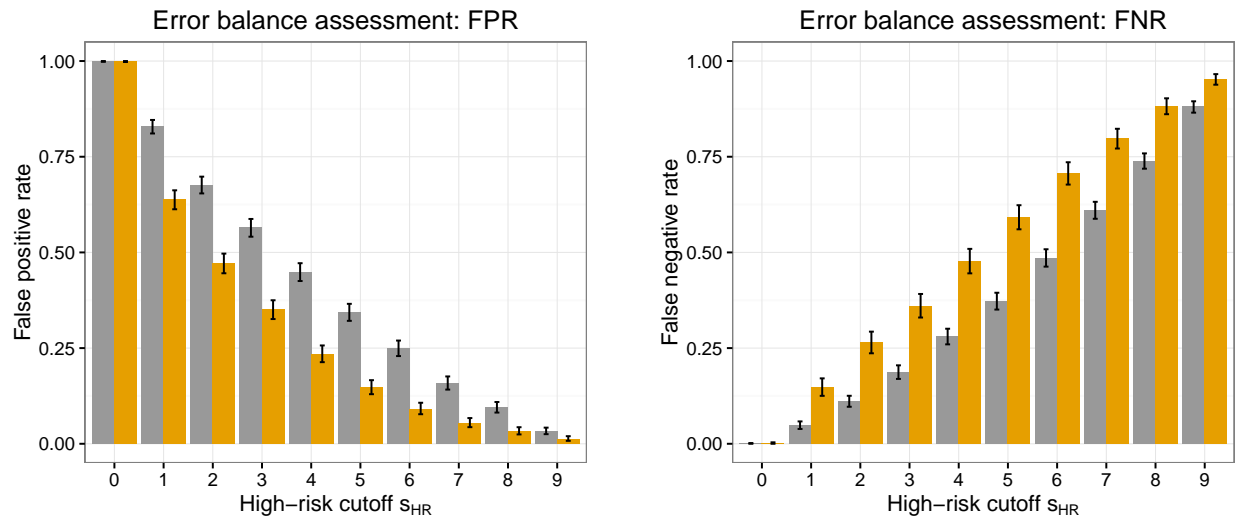
All of the fairness metrics presented in Section 2.1 can be thought of as imposing constraints on

the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to



(a) Bars represent empirical estimates of the expressions in (2.1):  $\mathbb{P}(Y = 1 \mid S = s, R = r)$  for decile scores  $s \in \{1, \dots, 10\}$ .

(b) Bars represent empirical estimates of the expressions in (2.2):  $\mathbb{P}(Y = 1 \mid S > s_{HR}, R = r)$  for values of the high-risk cutoff  $s_{HR} \in \{0, \dots, 9\}$



(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3):  $\mathbb{P}(S > s_{HR} \mid Y = 0, R = r)$  for values of the high-risk cutoff  $s_{HR} \in \{0, \dots, 9\}$

(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4):  $\mathbb{P}(S \leq s_{HR} \mid Y = 1, R = r)$  for values of the high-risk cutoff  $s_{HR} \in \{0, \dots, 9\}$

Figure 1: Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

	Low-Risk	High-Risk
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Table 1: T/F denote True/False and N/P denote Negative/Positive. For instance, FP is the number of false positives: individuals who are classified as high-risk but who do not reoffend.

show that the prevalence ( $p$ ), positive predictive value (PPV), and false positive and negative error rates (FPR, FNR) are related via the equation

$$\text{FPR} = \frac{p}{1-p} \frac{1-\text{PPV}}{\text{PPV}}(1-\text{FNR}). \quad (2.6)$$

From this simple expression we can see that if an instrument satisfies predictive parity—that is, if the PPV is the same across groups—but the prevalence differs between groups, the instrument cannot achieve equal false positive and false negative rates across those groups.

This observation enables us to better understand why we observe such large discrepancies in FPR and FNR between black and white defendants in Figure 1. The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants. Thus at any threshold  $s_{\text{HR}}$  where the COMPAS RPI satisfies predictive parity, equation (2.6) tells us that some level of imbalance in the error rates must exist. Since not all of the fairness criteria can be satisfied at the same time, it becomes important to understand the potential impact of failing to satisfy particular criteria. This question is explored in the context of a hypothetical risk-based sentencing framework in the next section.

### 3 Assessing impact

In this section we show how differences in false positive and false negative rates can result in disparate impact under policies where a high-risk assessment results in a stricter penalty for the defendant. Such situations may arise when risk assessments are used to inform bail, parole, or sentencing decisions. In Pennsylvania and Virginia, for instance, statutes permit the use of RPI’s in sentencing, provided that the sentence ultimately falls within accepted guidelines[1]. We use the term “penalty” somewhat loosely in this discussion to refer to outcomes both in the pre-trial and post-conviction phase of legal proceedings. For instance, even though pre-trial outcomes such as the amount at which bail is set are not punitive in a legal sense, we nevertheless refer to bail amount as a “penalty” for the purpose of our discussion.

There are notable cases where RPI’s are used for the express purpose of informing risk reduction efforts. In such settings, individuals assessed as high risk receive what may be viewed as a benefit rather than a penalty. The PCRA score, for instance, is intended to support precisely this type of decision-making at the federal courts level [11]. Our analysis in this section specifically addresses use cases where high-risk individuals receive stricter penalties.

To begin, consider a setting in which guidelines indicate that a defendant is to receive a penalty

$t_{\min} \leq T \leq t_{\max}$ . A very simple risk-based approach, which we will refer to as the MinMax<sup>a</sup> policy, would be to assign penalties as follows:

$$T_{\text{MinMax}}(s) = \begin{cases} t_{\min} & \text{if } s > s_{\text{HR}} \\ t_{\max} & \text{if } s < s_{\text{HR}} \end{cases}. \quad (3.1)$$

In this simple setting, we can precisely characterize the extent of disparate impact in terms of recognizable quantities. Our analysis will focus on the quantity

$$\Delta = \Delta(y_1, y_2) \equiv \mathbb{E}(T \mid R = b, Y = y_1) - \mathbb{E}(T \mid R = w, Y = y_2),$$

which is the expected difference in sentence duration between defendants in different groups, with potentially different outcomes  $y_1, y_2 \in \{0, 1\}$ .  $\Delta$  is taken to serve as our the measure of disparate impact.

**Proposition 3.1.** *The expected difference in penalty under the MinMax policy is given by*

$$\begin{aligned} \Delta &\equiv \mathbb{E}(T \mid R = b, Y = y_1) - \mathbb{E}(T \mid R = w, Y = y_2) \\ &= (t_{\max} - t_{\min})(\mathbb{P}(S > s_{\text{HR}} \mid R = b, Y = y_1) - \mathbb{P}(S > s_{\text{HR}} \mid R = w, Y = y_2)) \end{aligned}$$

A proof can be found in Appendix A. We will discuss two immediate Corollaries of this result.

**Corollary 3.1** (Non-Recidivists). *Among individuals who do not recidivate, the difference in average penalty under the MinMax policy is*

$$\Delta = (t_{\max} - t_{\min})(\text{FPR}_b - \text{FPR}_w), \quad (3.2)$$

where  $\text{FPR}_r$  denotes the false positive rate among individuals in group  $R = r$ .

**Corollary 3.2** (Recidivists). *Among individuals who recidivate, the difference in average penalty under the MinMax policy is*

$$\Delta = (t_{\max} - t_{\min})(\text{FNR}_w - \text{FNR}_b), \quad (3.3)$$

where  $\text{FNR}_r$  denotes the false negative rate among individuals in group  $R = r$ .

When using an RPI that satisfies predictive parity in populations where recidivism prevalence differs across groups, it will generally be the case that the higher recidivism prevalence group will have a higher FPR and lower FNR. From equations (3.2) and (3.3), we can see that this would on average result in greater penalties for defendants in the higher prevalence group, both among recidivists and non-recidivists.

An interesting special case to consider is one where  $t_{\min} = 0$ . This could arise in sentencing decisions for offenders convicted of low-severity crimes who have good prior records. In such cases, so-called restorative sanctions may be imposed as an alternative to a period of incarceration. If

---

<sup>a</sup>The term MinMax as used throughout this paper has no intended connection the decision-theoretic notion of minimax decision rules. Min and Max in this context refer to the minimum and maximum allowable sentences as stipulated by sentencing guidelines.

we further take  $t_{\max} = 1$ , then  $\mathbb{E}T = \mathbb{P}(T \neq 0)$ , which can be interpreted as the probability that a defendant receives a sentence imposing some period of incarceration.

It is easy to see that in such settings a non-recidivist in group  $b$  is  $\text{FPR}_b/\text{FPR}_w$  times more likely to be incarcerated compared to a non-recidivist in group  $w$ .<sup>a</sup> This naturally raises the question of whether overall differences in error rates are observed to persist across more granular subpopulations, such as the subset of individuals eligible for restorative sanctions. We explore this question in the section below.

### 3.1 Conditioning on other covariates

One might expect that differences in false positive rates are largely attributable to the subset of defendants who are charged with more serious offenses and who have a larger number of prior arrests/convictions. While it is true that the false positive rates within both racial groups are higher for defendants with worse criminal histories, considerable between-group differences in these error rates persist across low prior count subgroups. Figure 2 shows plots of false positive rates across different ranges of prior count for all defendants and also for the subset charged with a misdemeanor offense, which is the lowest severity criminal offense category. As one can see, differences in false positive rates between Black defendants and White defendants persist across prior record subgroups.

In general, all of the theoretical results presented in this section extend to the setting where we further condition on the covariates  $X$ . The main difference is that all classification metrics would need to be evaluated conditional on  $X$ . For instance, assuming that  $t_{\min}$  and  $t_{\max}$  are constant on a set  $\mathcal{X}$ , Corollary 3.1 would say that the difference in average penalty under the MinMax policy among non-recidivists for whom  $X \in \mathcal{X}$  is given by

$$\Delta = (t_{\max} - t_{\min}) (\text{FPR}_b(\mathcal{X}) - \text{FPR}_w(\mathcal{X})) \tag{3.4}$$

$$\equiv (t_{\max} - t_{\min}) (\mathbb{P}(S > s_{\text{HR}} \mid R = b, Y = 0, X \in \mathcal{X}) - \mathbb{P}(S > s_{\text{HR}} \mid R = w, Y = 0, X \in \mathcal{X})). \tag{3.5}$$

The false positive rates shown in Figure 2(a) correspond precisely to the quantities  $\text{FPR}_r(\mathcal{X})$  for choices of  $\mathcal{X}$  given by different prior record count bins. The leftmost bars correspond to taking  $\mathcal{X} = \{\#\text{priors} = 0\}$ . Similarly the leftmost bars in Figure 2(b) correspond to taking  $\mathcal{X} = \{\#\text{priors} = 0, \text{charge degree} = M\}$ . In Appendix B we present a logistic regression analysis showing that significant differences in false positive rates persist even after adjusting for a number of other recidivism-related covariates.

### 3.2 Connections to measures of differences in distribution

In their analysis of the PCRA instrument, Skeem and Lowenkamp [11] remark that some applications of the risk score could create disparate impact due to differences in the score distributions between black and white offenders. To summarize the distributional difference in scores between the two groups, the authors report a Cohen’s  $d$  of 0.34, with a corresponding non-overlap of 13.5%.

---

<sup>a</sup>We are overloading notation in this expression: Here,  $\text{FPR}_r = \mathbb{P}(\text{HR} \mid R = r, t_L = 0)$ , similarly for  $\text{FNR}_r$ .

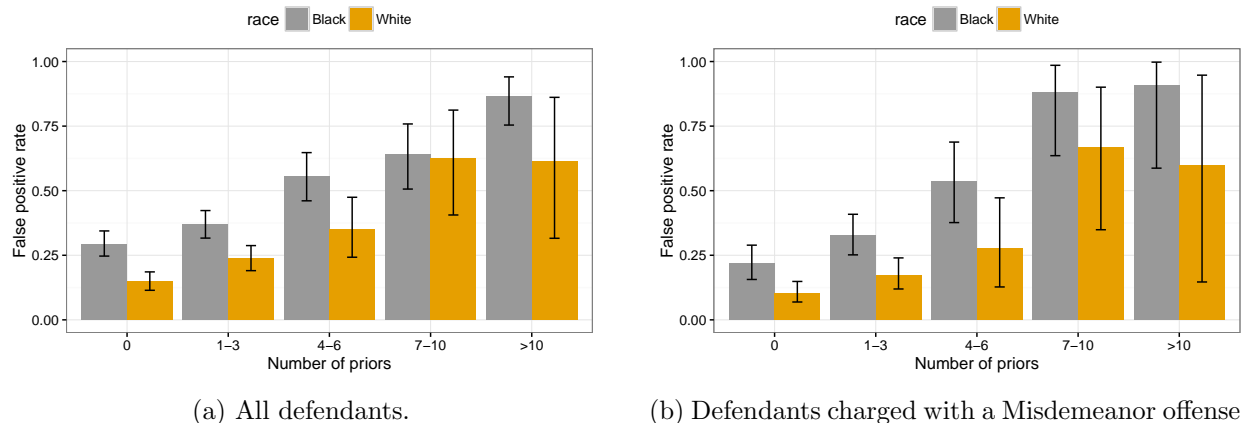


Figure 2: False positive rates across prior record count. Plot is based on assessing a defendant as “high-risk” if their COMPAS decile score is  $> s_{HR} = 4$ . Error bars represent 95% confidence intervals.

A natural question to ask is whether the level of disparity in sentence duration,  $\Delta$ , is in some sense closely related to such measures of distributional difference. With a small generalization of the % *non-overlap* measure, we can answer this question in the affirmative.

The % non-overlap of two distributions is generally calculated assuming both distributions are normal, and thus has a one-to-one correspondence to Cohen’s  $d$  [27].<sup>a</sup> However, as we can see from Figure 3, the COMPAS decile score is far from being normally distributed in either group. A more reasonable way to calculate % non-overlap in such cases is to note that in the Gaussian case % non-overlap is equivalent to the total variation distance. Letting  $f_{r,y}(s)$  denote the score distribution among individuals in group  $r$  with recidivism outcome  $y$ , one can establish the following sharp bound on  $\Delta$ .

**Proposition 3.2** (Percent overlap bound). *Under the MinMax policy,*

$$\Delta(y_1, y_2) \leq (t_{\max} - t_{\min})d_{TV}(f_{b,y_1}, f_{w,y_2}).$$

This result is simple to understand. When there is some non-overlap between the score distributions for two groups, the worst case scenario is that the non-overlap is entirely due to mass shifting from scores below  $s_{HR}$  to those above  $s_{HR}$ . In such cases, the inequality becomes an equality.

### 3.3 Empirical results

In this section we present some empirical results based on two hypothetical sentencing rules: the *MinMax* rule introduced in the previous section, and the *Interpolation* rule, which we will introduce below. Though the offenders in our data set come from Broward County, Florida, our empirical analysis is modelled on the sentencing guidelines of the State of Pennsylvania.

<sup>a</sup> $d = \frac{\bar{S}_b - \bar{S}_w}{SD}$ , where  $SD$  is a pooled estimate of standard deviation.



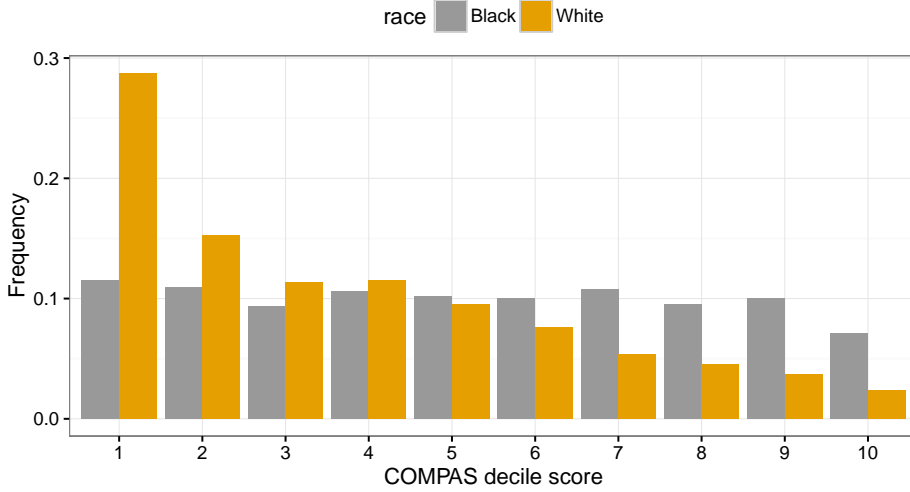


Figure 3: COMPAS decile score histograms for Black and White defendants. Cohen’s  $d = 0.60$ , non-overlap  $d_{TV}(f_b, f_w) = 24.5\%$ .

The penalty ranges  $t_{\min}$  and  $t_{\max}$  are selected by approximately matching each offender’s charge degree (M2 - F1) to a sentence range in Pennsylvania’s Basic Sentencing Matrix (PA Code §303.16). This matrix provides sentence ranges based on the charge degree for the current offense and the defendant’s prior record score (0 - 5+). We do not have enough information in the Broward County data to reliably assign a prior record score for each individual. Our results are based on using the sentencing range corresponding to a prior record score of 1 for all defendants in the data.

Figure 4 shows the expected sentences for black and white defendants broken down by observed recidivism outcome. The  $x$ -axis in these figures is taken to be the offense gravity score, which for the purpose of this analysis is mapped to charge degree as indicated in Table 2.

Offense gravity score	2	3	5	7	8
Charge Degree	(M2)	(M1)	(F3)	(F2)	(F1)

Table 2: Mapping between offense gravity score and charge degree used in the empirical analysis.

Results are shown for both the MinMax policy introduced earlier in this section, and the Interpolation policy, which is given by

$$T_{\text{Int}}(s) = t_{\min} + \frac{s - 1}{9}(t_{\max} - t_{\min}). \tag{3.6}$$

Unlike the MinMax policy, which is based on the coarsened score, the Interpolation policy assigns sentences by linearly interpolating between  $t_{\min}$  and  $t_{\max}$  based on the assigned decile score. We see that under both policies there are consistent trends in the expected sentences. Black defendants are observed to receive higher sentences than white defendants both within the non-recidivating subgroup and the recidivating subgroup (except in the F1 charge degree category, where sample sizes are small and results are non-significant). Since white defendants have higher false negative rates and lower false positive rates than black defendants, the empirical results are consistent with the theoretical results presented earlier in this section.

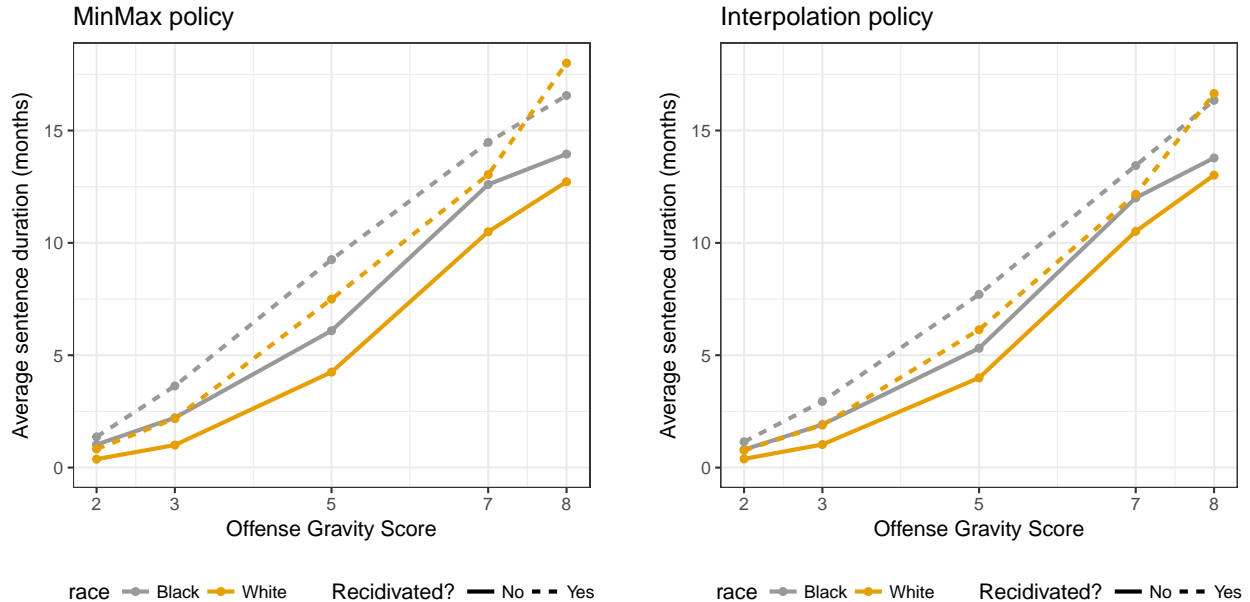


Figure 4: Average sentences under the hypothetical sentencing policies described in Section 3.3. The mapping between the  $x$ -axis variable and the offender’s charge degree is given in Table 2. For all OGS levels except 8, observed differences in average sentence are statistically significant at the 0.01 level.

## 4 Revisiting predictive parity

In this final section we revisit the notion of predictive parity and further discuss its implications for general classifiers. We know from equation (2.6) that when the positive predictive values are constrained to be equal but the prevalences differ across groups, the false positive and false negative rates cannot both be equal across those groups. While we have no direct control over recidivism prevalence, we do have some control over the PPV and error rates of our classifiers. At least in principle, we are free to tune our classifiers in any of the following ways:

- (i) Allow unequal false negative rates to retain equal PPV’s and achieve equal false positive rates
- (ii) Allow unequal false positive rates to retain equal PPV’s and achieve equal false negative rates
- (iii) Allow unequal PPV’s to achieve equal false positive and false negative rates

Figure 5 helps to put these trade-offs into perspective. From (2.6), we can see that FPR is a linear function of FNR under constraints on PPV and  $p$ . This means that, if PPV is fixed at a given value, tuning strategy (i) may require a very large increase in FNR in order to balance FPR. The black line shows feasible combinations of  $(FNR_b, FPR_b)$  when  $PPV_b$  is forced to equal the observed value  $PPV_w = 0.591$ . We can see that to get  $FPR_b$  to match  $FPR_w$ , we would need to increase  $FNR_b$  to around 0.7, which would be a substantial drop in accuracy. In view of Corollaries 3.1 and 3.2 Strategies (i) and (ii) may generally be undesirable because while they reduce disparate impact for one subgroup (e.g., among non-recidivists), they may increase it in the other.

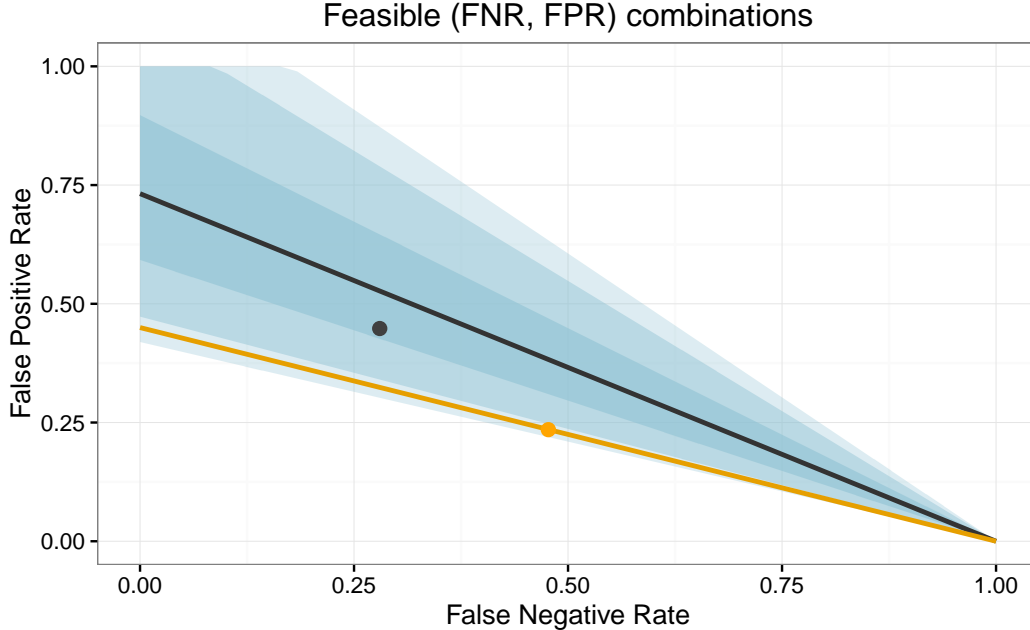


Figure 5: The two points represent the observed values of (FNR, FPR) for Black and White defendants. The orange line represents feasible values of (FNR, FPR) for White defendants when the prevalence  $p_w$  and  $PPV_w$  are both held fixed at their observed values in Table 1. The dark grey line represents feasible values of  $(FNR_b, FPR_b)$  when the prevalence  $p_b$  is held fixed at the observed value and  $PPV_b$  is set equal to the observed value of  $PPV_w = 0.591$ . *Nested* shaded regions correspond to feasible values of  $(FNR_b, FPR_b)$  if we allow  $PPV_b$  to vary under the constraint  $|PPV_b - 0.591| < \delta$ , with  $\delta \in \{0.05, 0.1, 0.125\}$ . The smaller  $\delta$ , the smaller the feasible region.

The preferred approach, at least in some cases, may be to pursue strategy (iii). This amounts to using a score  $S$  that does not satisfy predictive parity in the first place, but can also be achieved by allowing the high-risk cutoff  $s_{HR,r}$  to differ across groups. The shaded regions in Figure 5 show feasible values of  $(FNR_b, FPR_b)$  when we allow  $PPV_b$  to be within some  $\delta$  of the observed value of  $PPV_w$ . We can see that even at small values of  $\delta$  the feasible region is quite large.

## 5 Discussion

The primary contribution of this paper was to show how disparate impact can result from the use of a recidivism prediction instrument that is known to satisfy the fairness criterion of predictive parity. Our analysis focussed on the simple setting where a binary risk assessment was used to inform a binary penalty policy. While all of the formulas have natural analogs in the non-binary score and penalty setting, we find that many of the salient features are already present in the analysis of the simpler binary-binary problem.

A key limitation of our analysis stems from potential biases in the observed data that may affect our ability to draw valid inferences concerning the fairness of an RPI. Throughout this paper we have implicitly operated under the assumption that the observed recidivism outcome  $Y$  is a suitable outcome measure for the purpose of assessing the fairness properties of a recidivism

prediction instrument. However, the true outcome of interest in this context is *reoffense*, which is not what we observe. In the latest statistics released by the Federal Bureau of Investigation[28], it is reported that 46% of violent crimes and 19.4% of property crimes were successfully cleared by law enforcement agencies. Many criminal offenders are simply never identified. It is therefore possible that a non-negligible fraction of the individuals in our data for whom we observed  $Y = 0$  did in truth reoffend. If this is indeed the case, and if there are group differences in the rates at which offenders are caught, the findings of empirical fairness assessments may be misleading. Understanding how such forms of data bias affect the ability to assess instruments with respect to different fairness criteria is a subject of our ongoing research efforts.

## 6 Conclusion

In closing, we would like to note that there is a large body of literature showing that data-driven risk assessment instruments tend to be more accurate than professional human judgements [29, 30], and investigating whether human-driven decisions are themselves prone to exhibiting racial bias [31, 32]. We should not abandon the data-driven approach on the basis of negative headlines. Rather, we need to work to ensure that the instruments we use are demonstrably free from the kinds of biases that could lead to disparate impact in the specific contexts in which they are to be applied.

## A Proofs

*Proof of Proposition 3.1.* To simplify notation, we let  $HR$  denote the event  $\{S > s_{HR}\}$ .

$$\begin{aligned}
\mathbb{E}(\Delta(y_1, y_2)) &= \mathbb{E}(T \mid R = b, Y = y_1) - \mathbb{E}(T \mid R = w, Y = y_2) \\
&= t_{\max} \mathbb{P}(HR \mid R = b, Y = y_1) + t_{\min}(1 - \mathbb{P}(HR \mid R = b, Y = y_1)) \\
&\quad - t_{\max} \mathbb{P}(HR \mid R = w, Y = y_2) - t_{\min}(1 - \mathbb{P}(HR \mid R = w, Y = y_2)) \\
&= t_{\max}(\mathbb{P}(HR \mid R = b, Y = y_1) - \mathbb{P}(HR \mid R = w, Y = y_2)) \\
&\quad + t_{\min}(\mathbb{P}(HR \mid R = w, Y = y_2) - \mathbb{P}(HR \mid R = b, Y = y_1)) \\
&= (t_{\max} - t_{\min})(\mathbb{P}(HR \mid R = b, Y = y_1) - \mathbb{P}(HR \mid R = w, Y = y_2))
\end{aligned}$$

□

*Proof of Proposition 3.2.* By definition of total variation distance, for any event  $A$ ,

$$|\mathbb{P}(A \mid R = b, Y = y_1) - \mathbb{P}(A \mid R = w, Y = y_2)| \leq d_{TV}(f_{b,y_1}, f_{w,y_2})$$

Applying this inequality to Proposition 3.1 with  $A = \{S_c = HR\}$  gives

$$\begin{aligned}
\mathbb{E}(\Delta(y_1, y_2)) &= (t_{\max} - t_{\min})(\mathbb{P}(HR \mid R = b, Y = y_1) - \mathbb{P}(HR \mid R = w, Y = y_2)) \\
&\leq (t_{\max} - t_{\min})d_{TV}(f_{b,y_1}, f_{w,y_2})
\end{aligned}$$

□

## B Covariate-adjusted false positive rates

In this section we present the results of a logistic regression analysis that we conducted in order to assess whether the observed differences in false positive rates between black and white defendants can be entirely accounted for by other covariates. We find that adjusting for covariates decreases the gap, but it nevertheless remains large and statistically significant.

For the purpose of this analysis we consider only the subset of defendants who *do not* recidivate. The outcome variable for the logistic regression is taken to be

$$y = \begin{cases} 1, & S > 4 \\ 0, & S \leq 4 \end{cases},$$

where  $S$  denotes the COMPAS decile score. In this setup,  $y = 0$  denotes a True Negative and  $y = 1$  denotes a False Positive. Statistically significant positive coefficient estimates correspond to variables associated with increased likelihood of false positives.

Table 3 shows the results of regressing  $y$  on race alone. The coefficient of race in this model is large, positive, and statistically significant. Without adjusting for other covariates, the odds that a non-recidivating Black defendant receives a high-risk assessment are  $e^{0.976} = 2.6$  times higher than those of a White defendant.

Table 4 shows the results of regressing  $y$  on race, age, gender, number of priors, and charge degree. The coefficient of race is smaller than it was in the un-adjusted model, but it is nevertheless large and statistically significant. Even after adjusting for these other factors, the odds that a non-recidivating Black defendant receives a high-risk assessment are  $e^{0.547} = 1.72$  times higher than those of a White defendant.

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.183	0.061	-19.33	0.0000
raceBlack	0.976	0.077	12.60	0.0000

Table 3: Logistic regression with race alone.

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.397	0.176	7.92	0.0000
raceBlack	<b>0.547</b>	0.087	6.30	0.0000
Age	-0.079	0.005	-17.48	0.0000
sexMale	-0.291	0.098	-2.97	0.0030
Number of Priors	0.283	0.016	17.78	0.0000
chargeMisdemeanor	-0.109	0.088	-1.25	0.2123

Table 4: Logistic regression with race and other covariates that may be associated with recidivism

## References

- [1] Model penal code: Sentencing. American Law Institute, 2016.
- [2] Thomas Blomberg, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. Validation of the compas risk assessment classification instrument. 2010.
- [3] Ben Casselman Anna Maria Barry-Jester and Dana Goldstein. Should prison sentences be based on crimes that haven’t been committed yet?
- [4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks. 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [5] Northpointe. Compas risk & need assessment system: Selected questions posed by inquiring agencies.
- [6] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to “machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks.”. *Unpublished manuscript*, 2016.
- [7] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.
- [8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. How we analyzed the compas recidivism algorithm. 2016. URL <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- [9] Administrative Office of the United States Courts. An overview of the federal post conviction risk assessment, September 2011.
- [10] Jay P Singh. Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, 31(1):8–22, 2013.
- [11] Jennifer L Skeem and Christopher T Lowenkamp. Risk, race, & recidivism: Predictive bias and disparate impact. *Available at SSRN*, 2015.
- [12] Jennifer L Skeem, John Monahan, and Christopher T Lowenkamp. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Available at SSRN 2718460*, 2016.
- [13] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.
- [14] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.
- [15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.
- [16] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

- [17] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.
- [18] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05):582–638, 2014.
- [19] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. 2016.
- [20] Kelly Hannah-Moffat. Actuarial sentencing: An ‘unsettled’ proposition. *Justice Quarterly*, 30(2):270–296, 2013.
- [21] Jennifer Skeem. Risk technology in sentencing: Testing the promises and perils (commentary on hannah-moffat, 2011). *Justice Quarterly*, 30(2):297–303, 2013.
- [22] John Monahan and Jennifer L Skeem. Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12:489–513, 2016.
- [23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
- [24] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it’s actually not that clear.
- [25] Nancy Ritter. Predicting recidivism risk: New tool in philadelphia shows great promise. *National Institute of Justice Journal*, 271, 2013.
- [26] PCS. Validation of risk scale. Technical report, Pennsylvania Commission on Sentencing, 2013.
- [27] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Lawrence Erlbaum Associates, 1988.
- [28] Uniform crime report: Crime in the united states, 2015 - offenses cleared. U.S. Department of Justice, 2016.
- [29] Paul E Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, 1954.
- [30] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19, 2000.
- [31] Shamena Anwar and Hanming Fang. Testing for racial prejudice in the parole board release process: Theory and evidence. Technical report, National Bureau of Economic Research, 2012.
- [32] Laura T Sweeney and Craig Haney. The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, 10(2):179–195, 1992.

# A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

---

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17, 2016

This past summer, a heated debate broke out about a tool used in courts across the country to help make bail and sentencing decisions. It's a controversy that touches on some of the big criminal justice questions facing our society. And it all turns on an algorithm.

The algorithm, called COMPAS, is used nationwide to decide whether defendants awaiting trial are too dangerous to be released on bail. In May, the investigative news organization ProPublica claimed that COMPAS is biased against black defendants. Northpointe, the Michigan-based company that created the tool, released its own report questioning ProPublica's analysis. ProPublica rebutted the rebuttal, academic researchers entered the fray, this newspaper's Wonkblog weighed in, and even the Wisconsin Supreme Court cited the controversy in its recent ruling that upheld the use of COMPAS in sentencing.

It's easy to get lost in the often technical back-and-forth between ProPublica and Northpointe, but at the heart of their disagreement is a subtle ethical question: What does it mean for an algorithm to be fair? Surprisingly, there is a mathematical limit to how fair any algorithm — or human decision-maker — can ever be.

## How do you define 'fair'?

The COMPAS tool assigns defendants scores from 1 to 10 that indicate how likely they are to reoffend based on more than 100 factors, including age, sex and criminal history. Notably, race is not used. These scores profoundly affect defendants' lives: defendants who are defined as medium or high risk, with scores of 5-10, are more likely to be detained while awaiting trial than are low-risk defendants, with scores of 1-4.



We reanalyzed data collected by ProPublica on about 5,000 defendants assigned COMPAS scores in Broward County, Fla. (See the end of the post, after our names, for more technical details on our analysis.) For these cases, we find that scores are highly predictive of reoffending. Defendants assigned the highest risk score reoffended at almost four times the rate as those assigned the lowest score (81 percent vs. 22 percent).

But are the scores fair?

Northpointe contends they are indeed fair because scores mean essentially the same thing regardless of the defendant's race. For example, among defendants who scored a seven on the COMPAS scale, 60 percent of white defendants reoffended, which is nearly identical to the 61 percent of black defendants who reoffended.

Consequently, Northpointe argues, when judges see a defendant's risk score, they need not consider the defendant's race when interpreting it. The plot below shows this approximate equality between white and black defendants holds for every one of Northpointe's 10 risk levels.

But ProPublica points out that among defendants *who ultimately did not reoffend*, blacks were more than twice as likely as whites to be classified as medium or high risk (42 percent vs. 22 percent). Even though these defendants did not go on to commit a crime, they are nonetheless subjected to harsher treatment by the courts. ProPublica argues that a fair algorithm cannot make these serious errors more frequently for one race group than for another.

### **You can't be fair in both ways at the same time**

Here's the problem: it's actually impossible for a risk score to satisfy both fairness criteria at the same time.

The figure below shows the number of black and white defendants in each of two aggregate risk categories — "low" and "medium or high" — along with the number of defendants within each category who went on to commit another crime.

The plot illustrates four points:

- Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race; this is Northpointe's definition of fairness.
- The overall recidivism rate for black defendants is higher than for white defendants (52 percent vs. 39 percent).
- Black defendants are more likely to be classified as medium or high risk (58 percent vs. 33 percent). While Northpointe's algorithm does not use race directly, many attributes that predict reoffending nonetheless vary by race. For example, black defendants are more likely to have prior arrests, and since prior arrests predict reoffending, the algorithm flags more black defendants as high risk even though it does not use race in the classification.

- Black defendants who don't reoffend are predicted to be riskier than white defendants who don't reoffend; this is ProPublica's criticism of the algorithm.

The key — but often overlooked — point is that the last two disparities in the list above are mathematically guaranteed given the first two observations.

If the recidivism rate for white and black defendants is the same within each risk category, and if black defendants have a higher overall recidivism rate, then a greater share of black defendants will be classified as high risk. And if a greater share of black defendants are classified as high risk, then, as the plot illustrates, a greater share of black defendants who do not reoffend will also be classified as high risk.

If Northpointe's definition of fairness holds, and if the recidivism rate for black defendants is higher than for whites, the imbalance ProPublica highlighted will always occur. (Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan explore this idea further in their recent [paper](#).)

### **What should we do?**

It's hard to call a rule equitable if it does not meet Northpointe's notion of fairness. A risk score of seven for black defendants should mean the same thing as a score of seven for white defendants. Imagine if that were not so, and we systematically assigned whites higher risk scores than equally risky black defendants with the goal of mitigating ProPublica's criticism. We would consider that a violation of the fundamental tenet of equal treatment.

But we should not disregard ProPublica's findings as an unfortunate but inevitable outcome. To the contrary, since classification errors here disproportionately affect black defendants, we have an obligation to explore alternative policies. For example, rather than using risk scores to determine which defendants must pay money bail, jurisdictions might consider [ending bail](#) requirements altogether — shifting to, say, electronic monitoring so that no one is unnecessarily jailed.

### **COMPAS may still be biased, but we can't tell.**

Northpointe has refused to disclose the details of its proprietary algorithm, making it impossible to fully assess the extent to which it may be unfair, however inadvertently. That's understandable: Northpointe needs to protect its bottom line. But it raises questions about relying on for-profit companies to develop risk assessment tools.

Moreover, rearrest, which the COMPAS algorithm is designed to predict, may be a biased measure of public safety. Because of heavier policing in predominantly black neighborhoods, or bias in the decision to make an arrest, blacks may be arrested more often than whites who commit the same offense.

Algorithms have the potential to dramatically improve the efficiency and equity of consequential decisions, but their use also prompts [complex ethical and scientific questions](#). The solution is not to eliminate statistical risk assessments. The problems we discuss apply equally to human decision-makers, and humans are additionally biased in ways that machines are not. We

must continue to investigate and debate these issues as algorithms play an increasingly prominent role in the criminal justice system.

*Sam Corbett-Davies and Emma Pierson are PhD students in the computer science department at Stanford University.*

*Avi Feller is an assistant professor in the Goldman School of Public Policy at the University of California at Berkeley.*

*Sharad Goel is an assistant professor in the department of management science and engineering at Stanford University.*

*Note on methods: ProPublica obtained records for nearly 12,000 defendants in Broward County, Fla., who were assigned a COMPAS score in 2013-2014. ProPublica then determined which defendants were charged with new crimes in the subsequent two years, and made this data set publicly available. We focused on the 5,278 cases involving defendants who are either white or black, and for which a full two years of recidivism information is available. We excluded Hispanic defendants from our analysis because there are not many in this data set. The COMPAS tool also rates defendants on about two dozen other dimensions of risk, including likelihood to commit a violent crime, but here we consider only the overall recidivism score.*

 **1 Comment**



# A Five-Level Risk and Needs System: Maximizing Assessment Results in Corrections through the Development of a Common Language

**R. Karl Hanson, PhD**

Public Safety Canada

**Guy Bourgon, PhD**

Public Safety Canada

**Robert J. McGrath, MA**

Vermont Department of Corrections;  
McGrath Psychological Services, PC

**Daryl Kroner, PhD**

Department of Criminology and Criminal Justice,  
Southern Illinois University

**David A. D'Amora, MS, LPC, CFC**

The Council of State Governments Justice Center

**Shenique S. Thomas, PhD**

The Council of State Governments Justice Center

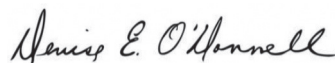
**Lahiz P. Tavaréz, BA**

The Council of State Governments Justice Center

Since 2009, the Bureau of Justice Assistance (BJA) has supported the National Reentry Resource Center (NRRC) to serve as the primary source of information and guidance in reentry, advancing the use of evidence-based practices and policies and creating a network of practitioners, researchers, and policymakers invested in reducing recidivism.

This white paper represents the culmination of two years of work undertaken as a special project of the NRRC. Initially aimed at improving the communication among justice practitioners and policymakers regarding risk information, a cornerstone of evidence-based practice, the collaborators made great advances over the course of the project, arriving at a thought-provoking framework for how to improve application of the Risk-Need-Responsivity (RNR) principles of evidence-based correctional intervention.

BJA is proud to have supported the development of this white paper, which we believe has the potential to improve justice system outcomes. Although much work remains—including pilot testing the model and tracking its impacts—we are hopeful that this paper can help move us all toward a “common language” of risk.



Denise O'Donnell, Director  
U.S. Department of Justice,  
Bureau of Justice Assistance

---

# Table of Contents

<b>Acknowledgments</b>	<b>1</b>
<b>Introduction</b>	<b>3</b>
<b>Expressing and Using Levels of Risk and Needs</b>	<b>6</b>
Statistical Indicators of Risk	6
Absolute Recidivism Rates	
Percentile Ranks	
Risk Ratios	
<b>A Five-Level Risk and Needs System</b>	<b>6</b>
Level I:	7
Correctional Response	
Prognosis	
Level II:	7
Correctional Response	
Prognosis	
Level III:	8
Correctional Response	
Prognosis	
Level IV:	8
Correctional Response	
Prognosis	
Level V:	9
Correctional Response	
Prognosis	
Returning to the Mr. Red Case Example	9
<b>Adopting the Five-Level Risk and Needs System</b>	<b>11</b>
<b>Conclusion</b>	<b>12</b>
<b>Appendix A. Table 1: Five-Level Risk and Needs System</b>	<b>13</b>
<b>Appendix B. Case Examples: Five Risk and Needs Levels</b>	<b>14</b>
<b>Key Terminology</b>	<b>17</b>
<b>Notes</b>	<b>18</b>

## Figures and Tables

<b>Figure 1.</b> The Number of People Expected to Reoffend Out of 100 in Each of the Five Standardized Risk and Needs Levels	10
<b>Table 1.</b> Five-Level Risk and Needs System	13

---

---

## Acknowledgments

This white paper was made possible through the support of the U.S. Department of Justice’s Bureau of Justice Assistance (BJA). In particular, we thank Associate Deputy Director Ruby Qazilbash and Senior Policy Advisor Juliene James at BJA for their leadership.

Thanks also go to the members of this publication’s group of expert advisors, listed below, who generously gave their time and expertise—providing insights and resources that were truly invaluable—and whose input greatly informed the development of the white paper.

This project was built on the strong foundational work done by the National Reentry Resource Center (NRRRC) in collaboration with partner researchers, risk and needs assessment instrument developers, practitioners, and leaders in the field. It draws on an extensive review of the literature and related research, observations of programs in the field, feedback from national experts, several multidisciplinary forums and advisory group discussions, and a rigorous review process. The following collaborators provided much-appreciated guidance and tireless support for this effort:

- Kelly Babchishin, University of Ottawa Institute of Mental Health Research; Karolinska Institutet
- John Baldwin, Iowa Department of Corrections
- Danielle Barron, Rhode Island Department of Corrections
- Jennifer Becan, Texas Christian University Institute of Behavioral Research
- David Berenson, Ohio Department of Rehabilitation and Correction
- Danica Binkley, Bureau of Justice Assistance
- Brandy Blasko, George Mason University
- Kevin Bowling, Global Strategic Solutions Working Group
- Tim Brennan, Northpointe Inc.
- Thurston Bryant, Bureau of Justice Assistance
- Bret Bucklen, Pennsylvania Department of Corrections
- Elizabeth Ann Carson, Bureau of Justice Statistics
- Brenda Crowding, California Department of Corrections and Rehabilitation
- Gary L. Dennis, Bureau of Justice Assistance
- Sarah Desmarais, North Carolina State University
- Amy Dezember, George Mason University
- Grant Duwe, Minnesota Department of Corrections
- Jennifer Ferguson, Maricopa County, Arizona, Adult Probation Department
- S.A. Godinez, Illinois Department of Corrections
- Zach Hamilton, Washington State University
- Jack Harne, National Institute of Justice
- Ed Hayes, Franklin County, Massachusetts, Sheriff’s Office
- Katie Herman, Center for Alternative Sentencing and Employment Services
- Daniel Heyns, Michigan Department of Corrections
- Beth Hoel, Maricopa County, Arizona, Adult Probation Department
- Seri Irazola, National Institute of Justice
- Mack Jenkins, San Diego County, California, Probation Department
- Deb Kerschner, Minnesota Department of Corrections
- KiDeuk Kim, the Urban Institute
- Doug Kosinski, Michigan Department of Corrections
- Ed Latessa, University of Cincinnati
- Chris Lobanov-Rostovsky, Colorado Sex Offender Management Board
- Zachary Lobel, U.S. House of Representatives
- Nathan Lowe, American Probation and Parole Association
- Chris Lowenkamp, Administrative Office of the U.S. Courts
- Audrey McAfee, Mississippi Department of Corrections



- 
- Karyn Milligan, San Diego County, California, Probation Department
  - Gary Mohr, Ohio Department of Rehabilitation and Correction
  - Ashley I. Moss, Office of U.S. Representative Hakeem Jeffries
  - Mary Ann Mowatt, American Probation and Parole Association
  - Katie Myers, Johnson County, Kansas, Department of Corrections
  - Janet Neeley, California Office of the Attorney General
  - Youlanda Nelson, Mississippi Department of Corrections
  - Mark Olver, University of Saskatchewan
  - Christina Ortiz-Marquez, Colorado Department of Corrections
  - Andrew Pallito, Vermont Department of Corrections
  - Lettie Prell, Iowa Department of Corrections
  - Rick Raemisch, Colorado Department of Corrections
  - Tom Roy, Minnesota Department of Corrections
  - Lee Seale, California Department of Corrections and Rehabilitation
  - Amy Seidlitz, Washington State Department of Corrections
  - Sharon Shipinski, Illinois Department of Corrections
  - Nate Simon, East Allegheny Supervision Unit, Pittsburgh, Pennsylvania, District Office
  - Jay P. Singh, Global Institute of Forensic Research
  - Paula Smith, University of Cincinnati Corrections Institute
  - Angeline Stanislaus, Missouri Department of Corrections and Sex Offender Rehabilitation and Treatment Services
  - Cynthia Stevens, Maricopa County, Arizona, Adult Probation Department
  - Raymond “Chip” Tafrate, Central Connecticut State University
  - Tony Tatman, Iowa Department of Corrections
  - Faye S. Taxman, George Mason University Center for Advancing Correctional Excellence
  - Gladyse Taylor, Illinois Department of Corrections
  - Donna Vittori, Maricopa County, Arizona, Adult Probation Department
  - Ashbel T. Wall, II, Rhode Island Department of Corrections
  - Bernard Warner, Washington State Department of Corrections
  - Angel Weant, Colorado State Judicial Department
  - John E. Wetzel, Pennsylvania Department of Corrections; The Council of State Governments Justice Center Board of Directors
  - Carl Wicklund, American Probation and Parole Association
  - Melanie Williams, Vermont Treatment Program for Sexual Abusers
  - Robin Wilson, Wilson Psychological Services LLC; McMaster University
  - Leonard Woodson, III, Colorado Department of Corrections, Sex Offender Treatment Program
  - Steve Wormith, University of Saskatchewan
- Special thanks go to Karen Watts and Bree Derrick at The Council of State Governments (CSG) Justice Center for their significant contributions throughout the development and drafting process, and to Michael D. Thompson and Suzanne Brown-McBride for their support of this project. The authors are also grateful to Katy Albis for her help producing this publication.
-

---

## Introduction

Risk and needs assessments are now routinely used in correctional systems in the United States to estimate a person’s likelihood of recidivism and provide direction concerning appropriate correctional interventions.<sup>1</sup> Specifically, they inform sentencing, determine the need for and nature of rehabilitation programs, inform decisions concerning conditional release, and allow community supervision officers to tailor conditions to a person’s specific strengths, skill deficits, and reintegration challenges. In short, risk and needs assessments provide a roadmap for effective correctional rehabilitation initiatives. When properly understood and implemented, they can help correctional organizations to provide the types and dosages of services that are empirically related to reductions in reoffending.<sup>2</sup>

Despite considerable advances in risk and needs assessment, however, the widespread use of a variety of risk and needs assessment instruments has created new challenges. Foremost, how do we compare the results of assessments conducted with different instruments? Although all of these instruments are trying to measure risk and needs, each instrument is unique in that it may comprise varying factors and weight those factors differently from other instruments. Furthermore, the field has not set standards or specifications about the terminology used to describe risk and needs categories across all of these instruments.<sup>3</sup> Although some risk and needs instruments use three nominal risk and needs categories (low, moderate, high), others use four nominal categories (low, low-moderate,

moderate-high, high), and still others use five (low, low-moderate, moderate, moderate-high, high). Some instruments use different terms entirely (e.g., poor, fair, good, very good).<sup>4</sup>

Complicating matters further, there are no standard definitions of these nominal risk and needs categories, so “low risk,” for example, might have different definitions from one instrument to the next. As such, the field of assessment and risk research struggles with perhaps its most significant obstacle: the absence of a precise, standardized language to communicate about risk. To further illustrate this problem, researchers<sup>5</sup> compared risk-level definitions among five assessment measures and found that only 3 percent of the people assessed were identified as high risk across all five *instruments* and only 4 percent of the people were identified as low risk by all five *measures*. This means that the same person can be described by different categories across different assessment instruments, or people in the same category can be described differently across different assessment instruments.

Beyond the lack of standard definitions of risk and needs categories, there is no consensus about what various labels mean with regard to the probability of reoffending or the specific profile of needs in each risk level.<sup>6</sup> This lack of consensus occurs not just across different instruments, but also across and within jurisdictions that use the same instrument but in different ways. The case study in Box 1 illustrates some of these challenges and the impact on the provision of effective correctional services.

### Box 1. Challenges of Applying Risk and Needs Assessments in Corrections: The Case of Mr. Red

Mr. Red was sentenced to prison for committing a violent offense while he was drunk. Prison staff assessed Mr. Red using their prison risk and needs assessment instrument and classified him as having a moderate level of risk and needs. This classification did not have much impact on the treatment services he received in prison, because every person in the prison with a history of committing a violent offense is referred to the same 24-hour anger management group and would not typically receive any other treatment services. Upon Mr. Red’s release from prison, his parole officer administered the parole risk and needs assessment instrument, which classified him as high risk. The parole officer talked with Mr. Red about what his score meant, which led them to work together to develop an individualized case plan. Commensurate with his high risk and needs classification, the initial plan included frequent contacts with parole staff, relatively restrictive supervision conditions, and a referral to an intensive substance use and cognitive skills program. A longer-term case plan included job training and possibly more treatment.

---

Mr. Red's case raises numerous questions:

How reliable and accurate were the two risk and needs assessment instruments administered to Mr. Red?

Assuming the two instruments were reliable and accurate, why might he be identified as different risk and needs levels by these two instruments?

Did the two instruments have the same number of risk and needs levels (e.g., three—low, moderate, and high) and how were these levels defined?

How was his case plan in prison and later in the community informed by his scores on the two risk and needs assessment instruments?

Were the differing amounts of treatment services Mr. Red received in prison and then in the community, along with the amount of community supervision and case management services he would receive, likely to increase, decrease, or have no effect on his risk of reoffending?

What is the appropriate level of treatment, supervision, and case management services for people who exhibit different levels of risk and needs?

How is a judge, probation or parole officer, treatment provider, or administrator to understand and communicate about what risk and needs assessment results mean?

For corrections and other criminal justice professionals, establishing a standard system for communicating about risk and needs levels would have tremendous benefits for the effectiveness of correctional systems. First, if professionals within and across jurisdictions used agreed-upon terms to describe risk and needs levels, everyone would have confidence that they knew what the terms meant, regardless of the instrument used. Consequently, they could have increased confidence that like people would be treated in like ways, regardless of the instrument used. Second, closely aligned, clearly defined, evidence-informed risk and needs levels would help to ensure that assessment results are used to determine the appropriate type and intensity of program and supervision resources and inform case planning. Third, this system would allow jurisdictions to save costs without jeopardizing public safety by more effectively matching interventions to people based on their likelihood of reoffending and their profile of needs and strengths. Fourth, for researchers, standardized risk and needs levels would facilitate comparative research, thereby further informing policy and practice.

Over the past two years, the NRRC, in partnership with Drs. Karl Hanson and Guy Bourgon of Public Safety Canada<sup>7</sup>, has facilitated efforts to examine and improve the standardization of the terminology associated with risk and needs levels and the

interpretation and application of risk and needs assessment results in correctional settings. From August 2014 to December 2015, the NRRC convened meetings of leading international experts on risk and needs assessments—including researchers from multiple disciplines, scientists, policymakers, and correctional practitioners—to develop a standard way to communicate about risk and needs, regardless of the assessment instrument in use.

This white paper reports the results of those efforts. It is written for researchers, practitioners, and policymakers who share the goal of reducing recidivism by improving the application of risk and needs assessments. Specifically, this white paper presents a model for supporting the implementation of Risk-Need-Responsivity (RNR) principles (see Box 2 on page 5)<sup>8</sup> through a standardized five-level risk and needs assessment system. The five levels are designed to inform case planning, guide how corrections and criminal justice professionals classify risk and needs, and help identify people who can benefit most from intervention. This empirically based system is intended to be broadly applicable and useful, and to increase the accountability of all system actors. Implementing this system does not require developing or adopting new risk and needs assessment instruments; rather, it involves realigning the existing information collected by agencies from

---

their validated risk and needs assessment instruments into a system that uses standard terminology. This standard terminology allows for greater clarity when people move from one part of the system to another or from one jurisdiction to another, facilitates clear communication between different treatment providers and correctional supervisors, and provides guidance

regarding treatment dosage and transition from one risk and needs level to another, regardless of what risk and needs assessment instrument is used or in what jurisdiction a person may reside.

## Box 2. Risk-Need-Responsivity (RNR) Principles

### The RNR principles have three major components:

**Risk Principle:** Match the intensity of services to a person’s level of risk for criminal activity

The risk principle states that the level of service should match a person’s risk of reoffending. Research shows that prioritizing supervision and program services for people at a moderate or higher risk of reoffending can lead to a significant reduction in recidivism for this population. Conversely, intensive interventions for people who are at a low risk of reoffending may actually be harmful and contribute to increasing the person’s likelihood of engaging in criminal behavior. High-intensity supervision or programming for lower-risk people has been shown to be an ineffective use of resources.

**Need Principle:** Target **criminogenic needs** (factors that contribute to the likelihood of new criminal activity)

The need principle directs that treatment and case management should prioritize the core criminogenic needs that can be positively impacted through services, supervision, and supports. Major criminogenic needs include **attitudes supportive of crime**, procriminal peers, lack of engagement in work/family, substance use, aimless use of leisure time, and **lifestyle instability**. Research indicates that the greater the number of criminogenic needs addressed through interventions, the greater positive impact those interventions will have on reducing recidivism.

**Responsivity Principle:** Account for a person’s abilities and learning styles when designing services

The responsivity principle highlights the importance of reducing barriers to learning by addressing learning style, reading ability, and motivation when designing supervision and program service strategies. The two types of responsivity—general and specific—have implications at the program and individual levels.

The general responsivity principle refers to the need for interventions that help to address criminogenic risk factors such as antisocial thinking. Research shows that social learning approaches and cognitive behavioral therapies can be effective in meeting a range of these needs, regardless of the type of crime committed. Prosocial modeling and skills development, teaching problem-solving skills, and using more positive than negative reinforcement have all been shown to be effective.

Specific responsivity refers to the principle that distinct personal needs should be addressed in order to prepare someone for receiving the interventions used to reduce criminal behavior. Specific responsivity relates to the “fine-tuning” of services or interventions, such as modifying a cognitive behavioral intervention to account for a cognitive impairment associated with mental illness. It also accounts for the person’s strengths; personality; learning style and capacity; motivation; and cultural, ethnic, racial, and gender characteristics, as well as behavioral health needs. Abiding by the responsivity principle can help to ensure that interventions are available and accessible and tailored to people in ways that can motivate them for services.

---

## Expressing and Using Levels of Risk and Needs

Assignment to a risk and needs level should have an empirical basis and be aligned with a recognizable pattern of meaningful, distinct characteristics. Useful risk and needs levels provide rich individual-level client information that includes statistical indicators about the likelihood of reoffending and the number and nature of risk-relevant propensities. These levels should also inform how one person compares with other people in the criminal justice system and inform appropriate correctional management strategies and treatment responses. There are several statistical indicators to use for describing people's risk and needs and for developing a common language to communicate this information.<sup>9</sup>

### Statistical Indicators of Risk

**Absolute Recidivism Rates.** An absolute recidivism rate is arguably the most useful and easily understood metric for reporting risk of reoffense. It is the percent likelihood of reoffending for people with the same risk score. Using the case illustration of Mr. Red (see Box 1, pages 3–4), an example of risk expressed as an absolute recidivism rate is the following: “Mr. Red’s score on the parole risk and needs assessment instrument was 42, which places him in the instrument’s high-risk category. People with scores in the high-risk category on the parole instrument have been found to have a 90 percent likelihood of being convicted of committing a new criminal offense within two years of returning to the community.” Risk and needs assessment instrument manuals include probability tables that report the actual or predicted reoffense rates linked to clusters of scores (i.e., nominal risk levels) or to each possible score on the assessment tool.

**Percentile Ranks.** Percentile ranks express the percentage of scores that are less than a given score. They are used to compare a person’s risk score with other people in the correctional population in a reference group, such as a representative sample from the person’s own jurisdiction. Options for comparing a person’s percentile rank to others include indicating that the person’s risk score (and risk of reoffending) is lower, the same, or higher in comparison to the reference group. The following is

an example of how percentile rank might be linked to nominal risk level: “Mr. Red’s score on the parole risk and needs assessment instrument places him in the top 5 percent in terms of risk to reoffend, so 95 percent of people in the reference group have a lower risk score than Mr. Red.” It can be advantageous to use percentile ranks because they are presented in a simple format and easily understood; however, they do not tell us what a person’s *actual* probability of reoffending is, or how it compares with others in the reference group.

**Risk Ratios.** Risk ratios show how a particular person’s risk to reoffend compares with that of the people who received an average score on the risk tool (i.e., the base rate of reoffending). There are several types of risk ratio statistics (e.g., rate ratio, hazard ratio, odds ratio). They vary from being complex to calculate and understand to being quite straightforward. Using a simple rate ratio statistic to add to what we already know about Mr. Red, we may say, “The risk of reoffending for people in Mr. Red’s category is two and half times higher than that of people who received an average score on the risk tool.” Simply put, if 40 out of 100 people reoffended over the course of 2 years, then the 2-year base rate of reoffending for that group of people is 40 percent. If Mr. Red’s relative risk were 2.5 times the base rate (2.5 times 40 percent equals 90 percent), then out of 100 high-risk people like Mr. Red, 90 would be expected to reoffend after 2 years.<sup>10</sup>

## A Five-Level Risk and Needs System

At the NRRC’s convening of risk and needs assessment advisors in August 2014, test developers and researchers considered what should be conveyed by nominal risk and needs levels and how many risk and needs levels are necessary to match people to appropriate supervision and services.<sup>11</sup> There was consensus that risk and needs assessment should go beyond simply categorizing people statistically. Rather, risk and needs assessment results should give us information about a person that will help guide *appropriate* and *differential* interventions and management strategies. Development of these strategies involves closely reviewing the domains captured in the risk and needs assessment. These

---

**domains** should include the underlying psychological, interpersonal, and lifestyle issues that relate to a person's criminogenic risk factors.<sup>12</sup> The *psychological domain* concerns cognitive, emotional, and behavioral features of a person that are empirically linked to offending. The *interpersonal domain* concerns a person's intimate, family, and peer relationships and how they support either prosocial or procriminal behavior. The *lifestyle domain* encompasses factors such as employment, education, housing, leisure activity, and substance use. When risk factors are grouped according to these risk- and needs-relevant domains in risk and needs assessment results, decision makers and treatment providers are positioned to understand the interconnection of a person's criminogenic needs, other life problems and circumstances, strengths, and likelihood of reoffending.

When considering the optimal number of risk and needs levels at the August 2014 convening, each member of the group presented a recommended number along with justification. Suggested options included from 2 to 11 levels, with serious consideration given to 3, 4, and 5 levels. Given our current knowledge of what works to reduce recidivism (e.g., providing treatment, supporting prosocial strengths, and the passage of time), there was sufficient evidence to support a five-level system. Subsequent field testing and consultations with program administrators, managers, analysts, and practitioners (i.e., clinicians) found that these five risk and needs levels, as described below, are highly recognizable to people working in corrections and align with many current practices. Field testing, however, generated little consensus on preferred names/labels for the levels. Consequently, the levels are labeled only by Roman numerals: I, II, III, IV, and V, with Level I describing the group of people identified with the lowest risk of reoffending and Level V describing the group of people with the highest risk of reoffending. Table 1 in Appendix A summarizes the five-level system.

## Level I

People assessed as Level I have few, if any, identifiable criminogenic or **non-criminogenic needs**. Any needs they exhibit are minimal and/or transitory in nature. Level I people have clearly identifiable resources and strengths within the psychological,

interpersonal, and lifestyle domains, and they are psychologically and socially similar to people without a criminal record. Their risk of new criminal behavior is no different from the rate of spontaneous, first-time offending for people without a criminal record, which is estimated at 1–2 percent per year among 18- to 25-year-old males,<sup>13</sup> with an upper limit of 5 percent over two years.

**Correctional Response.** Custody (i.e., placement in prison or jail) will be counterproductive in reducing recidivism for people grouped in Level I. The base rate of reoffending is low enough that prison may worsen recidivism outcomes.<sup>14</sup> People in this level are expected to comply with the conditions of community supervision, regardless of the supervision strategy, so minimal levels of monitoring would be warranted. The only human services needed are referral services and sharing of information on services and programs available in the community, such as family counseling.

**Prognosis.** The expected rate of reoffending for people in this level is very low.<sup>15</sup> Accordingly, there are not any expected changes in this level's base rate of reoffending because it is already low, and intervention is unlikely to lower it further. The risk of reoffending for this level is the same as the risk of criminal behavior for people in the community at large (less than or equal to 5 percent over three years).<sup>16</sup> The majority of people classified as Level I are expected to desist from criminal behavior, even without a correctional response.<sup>17</sup>

## Level II

People assessed as Level II have one or two identifiable criminogenic needs, and the severity of these needs is considered lower than the average risk defined in Level III. The needs are transitory or acute, rather than ingrained or sustained over time. People classified in Level II may have some non-criminogenic needs, but these, too, would not be severe. Like people assessed as Level I, Level II people have some identifiable resources and strengths. People in this level are expected to respond quickly and positively to services. The two-year rate of reoffending for this level is higher than for the community at large (i.e., greater than or equal to 5 percent), but is lower (estimated to be less than 30 percent) than the typical or average rate of reoffending for people designated as Level III (40 percent). The rate of reoffending for

---

this level is between 10 and 30 percent, with an average two-year reoffense rate of 19 percent.<sup>18</sup>

**Correctional Response.** Long-term custody of people identified as Level II would be counterproductive due to the negative effects of incarceration, such as destabilizing social supports and potentially increasing recidivism.<sup>19</sup> Members of this level are expected to comply with the conditions and requirements of community supervision. The most appropriate strategy for working with Level II people is simple, traditional case management to monitor compliance and service/program participation. In terms of human services, the focus should be on short-term interventions with an emphasis on problem solving and assistance in accessing community services.

**Prognosis.** By affording people identified as Level II with the correctional strategies outlined above, the majority would transition down to Level I and its respective rate of reoffending (i.e., less than 5 percent over two years) in a short time frame (e.g., six months or less).<sup>20</sup> Desistance from criminal offending is likely for those assessed as Level II when their criminogenic needs are addressed.

### Level III

Level III describes people in the middle of the risk and needs distribution of the entire correctional population (i.e., the national population of all people in custody or under community supervision). People identified as Level III have multiple criminogenic needs—varying in severity—in their psychological, interpersonal, and lifestyle domains. Generally, these people may have one or two discrete criminogenic needs that are considered primary drivers of their criminal behavior. People in Level III are also likely to have some non-criminogenic needs typical of the general correctional population (e.g., past trauma or mental health needs). Members of this level tend to have some identifiable resources and strengths, but their needs (criminogenic and non-criminogenic) are likely to be barriers to effective use of these resources and strengths. The rate of reoffending for Level III people who do not receive any interventions is equivalent to the overall correctional population's average rate of reoffending, presently estimated to be approximately 40 percent over two years.<sup>21</sup> The statistical boundaries of this risk level were designed to reflect the impact of routine effective correctional intervention, a reduction of approximately 10 percent in the absolute recidivism

rate.<sup>22</sup> Thus, using the 40 percent average reoffense rate, the upper boundary was set at about 10 percent higher (i.e., 49 percent) and the lower limit 10 percent lower (i.e., 30 percent).

**Correctional Response.** Custody for people grouped in Level III may be appropriate for short-term risk management. People in this level are expected to benefit from community supervision practices that both enhance compliance and encourage prosocial change. Human services should focus on the person's criminogenic needs, with secondary attention to non-criminogenic needs. The adequate dosage (i.e., duration and intensity) of services would amount to approximately 100–200 hours,<sup>23</sup> including formal treatment programs and change-focused supervision activities.

**Prognosis.** When people identified as Level III are provided with evidence-based correctional interventions in sufficient dosage, a significant reduction in reoffending would be expected—that is, a reduction of approximately 10 percent in the absolute recidivism rate.<sup>24</sup> Even when interventions are successful, however, the reoffense rate for Level III people would still be discernibly higher than the rate of offending for the population at large. For approximately half of people in Level III, successful interventions would result in reoffense rates that approximate the base rate of reoffending similar to that of people in Level II (i.e., 19 percent over two years). Nevertheless, it is expected that a proportion of these people would continue to be involved in the criminal justice system over the next three to five years, but over the longer term (five to seven years), desistance from crime would become increasingly likely.<sup>25</sup>

### Level IV

People assessed as Level IV have many criminogenic needs, likely representing all of the risk-relevant domains (psychological, interpersonal, and lifestyle), with a number of those needs being chronic and severe. In addition, these people have multiple, severe, and/or chronic non-criminogenic needs. The Level IV person may have some identifiable resources and strengths, but there are chronic barriers to accessing these resources, personal strengths, and social supports. The two-year rate of reoffending for people assessed as Level IV is approximately 65 percent (ranging from a low of 50 percent to a high of 84 percent), which is discernibly higher than the average

---

40-percent two-year rate of reoffending for the entire correctional population.

**Correctional Response.** The vast majority of people in Level IV have a history of incarceration, and when they are released to the community, they are likely to require intensive community supervision that is focused on monitoring for community safety, enhancing compliance, and strengthening treatment/service engagement, participation, and retention. Given the complexity and chronic nature of the criminogenic needs of people in this level, evidence indicates that intensive, lengthy (200–300 hours), and comprehensive services are required.<sup>26</sup> Correctional treatment and other social-service interventions should focus primarily on these people’s numerous criminogenic needs, and include services such as formal in-custody treatment programs, community-based treatment programs, and change-focused post-release supervision. Non-criminogenic needs should be addressed after Level IV people receive these services and begin to initiate prosocial lifestyle changes.

**Prognosis.** A significant reduction of reoffending (i.e., 10 percent) would be expected when people in Level IV are provided evidence-based correctional strategies in sufficient dosage. At best, however, the reoffense rate of these people would still be high, though reducing over time, and some of these people would show recidivism rates approximating those found in Level III. Given their chronic pattern of criminal behavior, the expectation is that a substantial proportion of Level IV people will reoffend over the long term, with a greater risk of recidivism sooner after release. Successful rehabilitation of these people typically involves gradual life changes over a long period of time (i.e., 10+ years) with increasingly lower rates of recidivism as they age.<sup>27</sup>

## Level V

People assessed as Level V have most, if not all, of the major criminogenic needs from the psychological, interpersonal, and lifestyle domains. Many of these needs are chronic, severe, and longstanding. In addition, these people likely have multiple, severe, and chronic non-criminogenic needs. Their identifiable resources and strengths are extremely limited, if they exist at all, or are used to support criminal behavior (e.g., superficial charm to support fraud). The base

rate of reoffending for Level V people (without intervention) is discernibly higher than that of Level IV. Their base rate of reoffending is that of people in the correctional population who reoffend most chronically (i.e., the highest 5 percent), with a corresponding minimum rate of reoffending of 85 percent within two years, and an average reoffense rate of approximately 90 percent.

**Correctional Response.** Custody is appropriate for people in Level V for the purposes of community safety. The degree of this group’s propensity to engage in criminal behavior warrants treatment services that are highly structured, comprehensive, intensive, and lengthy (e.g., well over 300 hours, provided over years). Ideally, the provision of services would occur within secure facilities prior to release, with gradual step-down of secure settings over time as the person demonstrates incremental behavioral change. People grouped in Level V are expected to require the most intensive community supervision, including close monitoring and surveillance as a priority for public protection. Change-focused supervision should gradually be introduced as the person demonstrates incremental behavioral and attitudinal change over time.

**Prognosis.** Reductions in reoffending for people in Level V take place gradually over decades, if at all.<sup>28</sup> Significant reductions of reoffending may be possible; however, evidence-based correctional strategies in sufficient dosage would be required. Nevertheless, their recidivism rates would be expected to remain high over the long term, eventually approaching the base rate of people grouped in Level IV after years of appropriate interventions. The chronic and persistent pattern of criminal behavior for people in Level V means that considerable time and intensive services would be required before they would be expected to approach the psychological profile and reoffending base rate of people grouped in Level III. In advanced age (50+), many could reach the reoffending base rate of Level II.<sup>29</sup>

## Returning to the Mr. Red Case Example

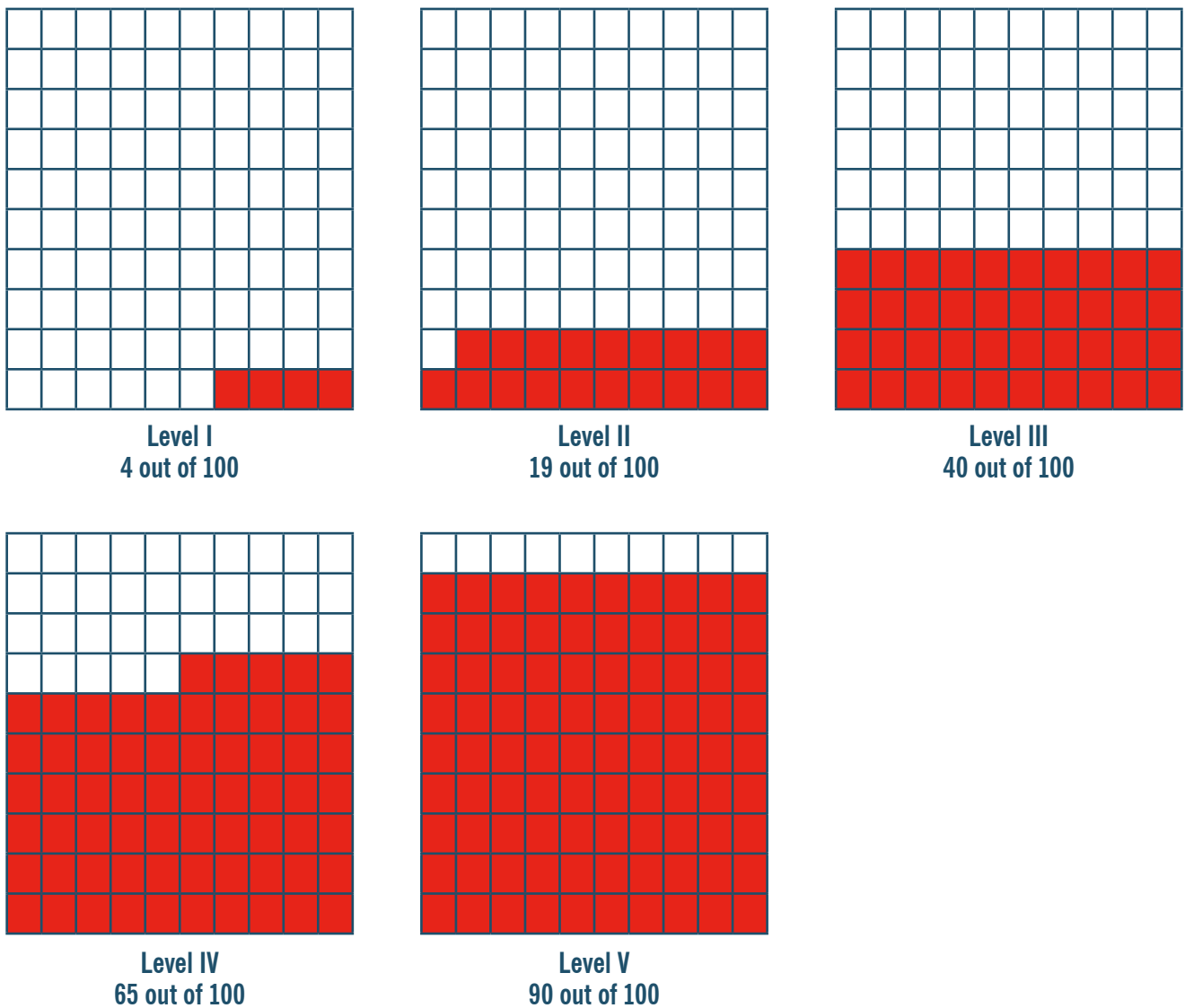
As shown in the Mr. Red case example, the five-level system allows us to identify and communicate about a person’s risk and needs using five groups and statistical indicators, such as percentile rank, risk ratios, and absolute recidivism. For a visual



representation of the five-level risk and needs assessment system we are proposing, consider Figure 1, which shows the expected reoffense rates of the five risk levels based on national samples,<sup>30</sup> with Level I representing the lowest risk and needs group and Level V the highest risk and needs group. Of the 100 small squares representing 100 people in each of the 5 boxes, those shaded red represent the expected number of people who will be convicted of a new offense within 2 years of placement in the community. Our fictional Mr. Red's risk and needs level would be

Level V. As the Level V graphic in Figure 1 shows, it is immediately evident that almost everyone (90 percent) assessed at this level of risk and needs is expected to reoffend within two years. Furthermore, by comparing the five levels, it is clear that the likelihood of reoffending of men similar to Mr. Red is significantly greater than that of people in the other four risk and needs levels. Appendix B provides more information about Mr. Red to illustrate psychological, interpersonal, and lifestyle domain factors relevant to his case, as well as his risk and needs assessment,

**Figure 1. The number of people expected to reoffend out of 100 in each of the five standardized risk and needs levels**  
(Red boxes indicate the number of people expected to reoffend.)



---

case recommendations, and prognosis. The appendix also contains four other case examples of the remaining risk and needs levels.

Each standardized risk and needs level in the system is associated with a certain number and severity of **dynamic risk factors** within the psychological, interpersonal, and lifestyle domains. The standardized five-level system informs correctional responses and offers prognoses of success. Without knowing anything about Mr. Red other than that he is in Level V, service providers and jurisdictions that adopt the five-level risk and needs system will immediately know important information about his risk, needs, recommended services, and prognosis (although the more comprehensive the risk and needs instrument that is used, the easier it is to identify and address the characteristics underlying a person's risk and needs). This system will be invaluable for communicating about and coordinating the delivery of his correctional services.

## Adopting the Five-Level Risk and Needs System

In order for the five-level system to be useful, test developers and jurisdictions must be able to adopt the system using their own instruments and data sets.<sup>31</sup> There are two separate issues that need be addressed before the five-level system can be used reliably. One, the recidivism boundaries that delineate the various levels require additional research to

further clarify the exact percentages. Specifically, the recidivism rate of the correctional population is presently estimated to be approximately 40 percent based on national statistics,<sup>32</sup> but these studies do not include descriptions of assessed risk levels and have some methodological limitations. We are cross-validating the "average" base-rate recidivism of three very large data sets to identify a more precise recidivism estimate. Regardless, the upper and lower recidivism boundaries of Level III are predetermined by the aforementioned treatment effect of a 10-percent reduction in recidivism, as are the defining characteristics of the five levels.

Further, in order for a jurisdiction to adopt the five-level risk and needs system, it must complete a validation study of its risk and needs assessment instrument that includes a sufficiently large and representative sample of people in the criminal justice system. A sufficient sample size is estimated to be approximately 500 people as long as the sample contains a minimum of 100 people who have reoffended within a follow-up period of two years. If the sample contains fewer than 100 people who have reoffended, then the sample size should be increased to meet this requirement.<sup>33</sup> Such a study permits the jurisdiction to (a) empirically demonstrate that the instrument it is using has at least moderate predictive accuracy (Area Under the Curve [AUC] values around .70); (b) establish reliable recidivism rates associated with each individual score of the risk and needs assessment instrument; and (c) identify the instrument's risk scores that are associated with each of the five levels.

### Box 3. Risk and Needs Assessment and Racial Disparity

Given the over-representation of people of color among those who are in the criminal justice system, it is important to consider how factors that influence decision making, including risk and needs assessment, can contribute to racial disparities in the justice system. Deliberate action should be taken to prevent racial bias from entering the risk and needs assessment process, including conducting a validation study whereby jurisdictions can confirm that the assessment instrument is accurate across all racial groups. Beyond validation, jurisdictions should have formal mechanisms in place to assess the quality of implementation of the risk and needs assessment instrument, and develop plans to address any bias found in the instrument itself or how it is being used. If used properly and effectively, risk and needs assessment can potentially help to limit racial bias in decision making in the criminal justice system by providing an objective, evidence-based assessment of criminogenic risk factors and needs.

---

The statistical analyses required to populate the categories are relatively simple. The assessment scores associated with reoffense rates of 5 percent or less after two years populate Level I. The scores of people whose reoffense rates are 85 percent or greater define Level V. The scores in the middle, Level III, represent people whose reoffense rates center on or are slightly above or below the average reoffense rate of the entire sample. In statistical terms, slightly above and slightly below are defined by the typical recidivism reduction observed in real-world implementation of cognitive-behavioral rehabilitation programs (i.e., an **r value** of .10 or 10 percent).<sup>34</sup> The scores for Levels II and IV are then quite simple to define: respectively, they are the remaining scores between Levels I and III and between Levels III and V.

In sum, Level III should be in the middle of the risk and needs distribution (centered on the median value of the risk tool). Level II should represent meaningfully lower risk and needs than average and Level IV should represent higher risk and needs than average. Those in Level I should have the same level of risk as the general population. People identified as Level V should have the very highest risk and needs. Their risk for recidivism is best managed through intensive community supervision and, in some cases, incarceration. Although the five-level risk and needs system was developed for general recidivism, the categories can also inform standardized risk category labels for other types of risk, such as sexual, spousal abuse, and any violent reoffending.<sup>35</sup>

---

## Conclusion

Over the past 20 years, there have been significant advances in understanding what works to reduce recidivism for people who have become involved with the criminal justice system. We know now that effective correctional intervention—meaning the implementation of evidence-based practices with fidelity—requires taking into account a person’s risk of reoffending and the needs that must be met to change that person’s behavior. Risk and needs assessments should inform case management, not just predict risk. Consequently, risk and needs assessments need to not only provide information concerning a person’s likelihood of reoffending but also identify that person’s needs and strengths to enable appropriate evidence-based correctional responses, and provide statistical data about the expected success of various

appropriate risk-reduction strategies. The five-level risk and needs system proposed in this paper synthesizes the empirical knowledge already captured by existing risk and needs assessment instruments, and it integrates what we know about effective correctional interventions, **life-course development**, and desistance from criminal behavior. Above all, the five risk and needs levels provide a system for criminal justice professionals to communicate about people precisely, clearly, and consistently, regardless of the jurisdiction where the assessment is conducted or the instrument that is used. By aligning correctional activities to these standardized levels, we increase the likelihood that people will actually receive the services and supervision they need to reduce recidivism.

# Appendix A

## Table 1: Five-Level Risk and Needs System

LEVEL	CRIMINOGENIC NEEDS	PROFILE AND 2-YEAR RECIDIVISM RATE WITHOUT INTERVENTION	SUPERVISION DOSE	CORRECTIONAL TREATMENT DOSE	TREATMENT EFFECT	PROGNOSIS FOLLOWING INTERVENTION
<b>I</b>	None or few – if any, mild and/or transitory	Non-offending profile: similar to people with no criminal record Average = 3% Range = less than 5%	Minimal or no monitoring	None – if needed, refer to community services	Risk so low that it will not be reduced further	Excellent, will stay in Level I
<b>II</b>	A few – some mild and transitory, or possibly acute	Vulnerable prosocial profile: higher risk than non-offending profile but lower than average Average = 19% Range = 5%–29%	Some – monitor for compliance, provide some change-focused interventions	Minimal – if any, very short term, refer to community services if needed	Risk so low that intervention can only have a minor impact	Very good, most will move from Level II to I
<b>III</b>	Multiple – some severe	Average offending profile: the middle of the risk and needs distribution Average = 40% Range = 30%–49%	Considerable – monitor for compliance and provide change-focused interventions	Significant – 100–200 hours	Intervention impact is significant and can meaningfully reduce reoffending	Good, many will move from Level III to II
<b>IV</b>	Multiple – some chronic and severe	Persistent offending profile: chronic and lengthy involvement in crime Average = 65% Range = 50%–84%	Intensive – monitor for safety and compliance, provide change-focused interventions	Very significant – 200–300 hours	Intervention impact can be significant but reduction will not quickly result in the lowest levels of risk	Improvement, some will move from Level IV to III, and as low as II after a significant period of time (i.e., 10+ years)
<b>V</b>	Multiple – chronic, severe, and entrenched, likely across psychological, interpersonal, and lifestyle domains	Entrenched criminal profile: virtually certain to reoffend Average = 90% Range = 85% or higher	Very intensive – monitor for safety and compliance, provide long-term and intensive change-focused interventions	Extensive – well over 300 hours, provided over years	Intervention can have an impact but initial risk so high that emphasis is on treatment readiness and behavioral management	Initial risk so high that reoffending will still be above average, some will move to Level IV or III, possibly as low as II in advanced age

---

## Appendix B

### Case Examples: Five Risk and Needs Levels

#### Level I

##### Mr. Green

**Background.** Mr. Green is 35 years old and was recently convicted of reckless driving causing injury. He was talking on his cell phone while driving his car, and he became distracted and hit a bicyclist, causing the person serious injury.

**Psychological Domain.** He has no previous history of criminal behavior or delinquency. He is thoughtful and goal oriented, expresses prosocial values, has remorse and takes responsibility for the crime, and is embarrassed by his actions.

**Interpersonal Domain.** He and his wife divorced seven years ago and share custody of their two children. They have a positive relationship. He has cohabited with his current girlfriend for the past three years. She is employed as a teacher and is prosocial. Their relationship appears quite healthy, and they socialize with prosocial peers from work, as well as the parents of his children's friends. His parents live nearby and they are prosocial and supportive.

**Lifestyle Domain.** He has worked full time since receiving his college degree about 14 years ago. He is a social drinker and has no history of drug abuse. He enjoys being involved in his children's activities, travels, plays in a basketball league, and plays cards with friends.

**Risk and Needs Assessment.** Mr. Green's score on the probation department's risk and needs assessment instrument was 3. This score identifies him as risk and needs Level I. Of 100 people with the same score, on average, 3 percent will be convicted of committing a new criminal offense within 2 years of placement in the community, with an upper limit of less than 5 percent.

**Recommendations and Prognosis.** Placement in prison or jail will be counterproductive in reducing recidivism for people in Level I, such as Mr. Green. The base reoffense rate is sufficiently low that prison may worsen recidivism outcomes. People in this level would be expected to comply with the conditions

of community supervision, regardless of the supervision strategy, so minimal levels of monitoring are warranted. The only human services that might be warranted would be sharing of information on and referral to services and programs available in the community.

#### Level II

##### Mr. Blue

**Background.** Mr. Blue is 32 years old and was recently convicted of driving while intoxicated (DWI) and possession of narcotics. At a routine traffic stop, he had a blood alcohol count of .10, and when his car was searched, police found him in possession of five grams of marijuana.

**Psychological Domain.** He successfully completed a year of probation following a conviction for assault at age 19. He is now more mature, and he is embarrassed about his offenses. He accepts responsibility for his actions and enrolled in alcohol treatment through his employee assistance program immediately after his arrest. He expresses prosocial values and respects authority.

**Interpersonal Domain.** He has been married for seven years, and he and his wife have two children. Their relationship is positive and stable. His wife works full time. They have several close friends, all of whom are employed and none have a criminal history. A few of his friends occasionally smoke marijuana. He and his wife are close to their families of origin, who are supportive and prosocial.

**Lifestyle Domain.** He has owned his own cleaning company for the past 5 years and employs 10 people. He has a history of "partying" as a teenager, and a recent assessment indicates alcohol use as "problematic" and drug use as "recreational." He is involved in many organized activities, including recreational hockey and a golf league.

**Risk and Needs Assessment.** Mr. Blue's score on the probation department's risk and needs assessment instrument was 15. This score identifies him as risk and needs Level II. Of 100 people with the same score, on average, 19 will be convicted of committing a new criminal offense within 2 years of placement in the community. Overall, the two-year recidivism rate of people in Level II ranges from 5 to 29 percent.

---

**Recommendations and Prognosis.** Long-term placement in prison or jail for people classified as Level II, such as Mr. Blue, would be counterproductive due to the negative effects of incarceration. Members of this level are expected to be compliant with the conditions and the requirements of community supervision. The most appropriate strategy for working with these people is simple, traditional case management to monitor compliance and service participation. In terms of human services, the focus should be on short-term interventions with an emphasis on social problem solving and assistance in obtaining existing community services. By affording people like Mr. Blue with appropriate short-term services, it is expected that they transition down to Level I within six months or less and that their rate of reoffending mirrors that of the general population.

### Level III Mr. Yellow

**Background.** Mr. Yellow is 32 years old and was recently convicted of DWI and driving without a license when he was stopped by police at 2 a.m. for erratic driving.

**Psychological Domain.** He has three previous convictions: one for a property offense in his early 20s, two DWIs in his mid-20s, and another DWI at age 30. He is generally prosocial. He views himself as a “blue-collar” man and does not identify himself as a criminal. He said he does not have a drinking problem and rationalizes his use of his vehicle without a license. He has poor problem-solving skills, is pessimistic about his life, has a rigid thinking style, and often makes impulsive decisions.

**Interpersonal Domain.** He has been divorced and now remarried for three years. He has one biological child and one stepchild. His relationship with his family is generally positive, with some discord about drinking and finances. He spends time primarily with coworkers in the construction trade and old friends, some of whom have criminal histories and most of whom drink. He has some interpersonal conflict with his boss at work. He has minimal contact with his father, who has a serious alcohol problem and was abusive. His mother passed away four years ago.

**Lifestyle Domain.** He has had fairly stable and full-time work with the same construction company for

the past four years, with sporadic seasonal layoffs. He typically arranges a short workday on Fridays and then meets his friends at a bar afterward. He has had an alcohol use problem for about 10 years and has never been in treatment. He is not involved in any organized leisure activities.

**Risk and Needs Assessment.** Mr. Yellow’s score on the probation department’s risk and needs assessment instrument was 24. This score identifies him as risk and needs Level III. His risk of reoffending is similar to that of people who receive an average score on the instrument. Of 100 people with the same score, on average, 40 will be convicted of committing a new criminal offense within 2 years of placement in the community. Overall, the two-year recidivism rate for people in Level III ranges from 30 to 49 percent.

**Recommendations and Prognosis.** People like Mr. Yellow should generally receive approximately 100–200 hours of formal treatment programming and change-focused supervision activities. If Mr. Yellow is given a jail or prison sentence, these interventions should be initiated while he is in custody. Level III people would be expected to benefit from treatment and community-supervision services that both enhance compliance and encourage prosocial change, and target criminogenic needs, with secondary attention to non-criminogenic needs. Services for people in Level III, compared with other levels, are likely to have the greatest impact on risk of reoffending. For approximately half of Level III people, successful intervention would result in reoffense rates similar to that of people in Level II (i.e., 19 percent over two years). Therefore, it is expected that a proportion of people in this level would continue to be involved in the criminal justice system over the next few (three to five) years, but over the longer term (five to seven years), desistance from crime would become increasingly likely.

### Level IV Mr. Orange

**Background.** Mr. Orange is 27 years old. He was recently convicted of committing three burglaries and possession of narcotics.

**Psychological Domain.** He has four previous criminal convictions in addition to a juvenile criminal history. He served two prior prison sentences for robbery, weapons possession, and drug-related offenses. He

---

committed his current offenses while on community supervision after he began using illegal drugs again. He has a history of problems with impulsivity and expresses procriminal and anti-authority values. His previous probation officer described him as likable and motivated to do well, but said that he leads a rather chaotic lifestyle.

**Interpersonal Domain.** He has a history of several short-term intimate relationships with women, most of whom have had substance use problems. He has no contact with his one child. His present partner of three months is prosocial and not a substance user, but most of his friends use illegal drugs. His only family contact is with a brother who regularly uses illegal drugs and has a lengthy criminal history.

**Lifestyle Domain.** He is presently unemployed, but typically works sporadically as a house painter during the building season. He has had a chronic alcohol and drug use problem since his teenage years. He regularly frequents local pubs, gambles through a bookie, and occasionally plays pickup basketball.

**Risk and Needs Assessment.** Mr. Orange's score on the probation department's risk and needs assessment instrument was 30. This score identifies him as risk and needs Level IV. Of 100 people with the same score, on average, 65 will be convicted of committing a new criminal offense within 2 years of placement in the community. Overall, the two-year recidivism rate of people in Level IV ranges from 50 to 84 percent.

**Recommendations and Prognosis.** Given the multiple, complex, and chronic nature of criminogenic needs among people grouped in Level IV, such as Mr. Orange, evidence indicates that intensive, lengthy (200–300 hours), and comprehensive treatment services are required to reduce reoffending. If Mr. Orange is given a jail or prison sentence, these treatment services should be initiated while he is in custody. When being supervised in the community, Level IV people would be expected to require intensive supervision, focusing on monitoring for community safety, enhancing compliance, and enhancing engagement in treatment and services. A significant reduction of reoffending (i.e., 10 percent) is expected when people like Mr. Orange receive evidence-based correctional programming in sufficient dosage. However, even when treatment is beneficial, the reoffending rate of these people would still be high, reducing only to the average reoffending rate

(the Level III base rate of 30 to 49 percent). Given the chronic pattern of criminal behavior, the expectation is that a substantial proportion of people in Level IV will reoffend over the long term. Successful rehabilitation for people in this level typically involves gradual life changes over a long period of time (i.e., 10+ years).

## Level V Mr. Red

**Background.** Mr. Red is 39 years old. His most recent convictions were for multiple counts of aggravated assault and kidnapping. Two incidents involved serious physical assaults on adult males, and one incident involved a woman whom he kidnapped and forced to withdraw money from an ATM. He successfully appealed legal errors made at his sentencing hearing, won early release from prison, and is now on probation. His earlier convictions include several property, drug, fraud, and violent offenses. He began getting in trouble with the law in early adolescence, has continued to engage in criminal behavior throughout his adulthood, and has a poor record of following community supervision conditions.

**Psychological Domain.** He presents to correctional staff as hostile and resentful of authority. He has a long history of acting impulsively. He values aggression and power as ways to get what he wants in life. He places blame on others for his own misdeeds and shows no remorse for his antisocial actions. He also shows pride in his long criminal history.

**Interpersonal Domain.** Although he has had many short-term sexual partners, he has never married or had long-term romantic relationships as an adult. He has been a gang member since his late teens. Many in his immediate family also have extensive criminal histories, and he has loose connections to most of them.

**Lifestyle Domain.** He often takes on the role of "enforcer" in his gang. He has little record of employment during the last several years. He has a lengthy history of drug and alcohol use, and he committed a significant portion of his offenses while under the influence of substances.

**Risk and Needs Assessment.** Mr. Red's score on the probation department's risk and needs assessment instrument was 42. This score identifies him as risk and needs Level V. Of 100 people with the

---

same score, approximately 90 will be convicted of committing a new criminal offense within 2 years of placement in the community. Overall, the two-year recidivism rate of people in Level V is 85 percent or greater.

**Recommendations and Prognosis.** Treatment services for people grouped in Level V, such as Mr. Red, need to be highly structured, comprehensive, intensive, and lengthy—well over 300 hours. If Mr. Red is sentenced to incarceration, these treatment

services should be initiated while he is in custody. If he is living in the community, intensive supervision with close monitoring and surveillance is a priority for public protection. People in Level V are described as participating in life-course persistent offending, meaning that considerable time and intensive services are required before they would be expected to benefit substantially from correctional intervention and reduce their risk to Level IV. In advanced age (50+), many could reach the reoffending base rate of Level III, which ranges from 30 to 49 percent.

---

## Key Terminology

**attitudes supportive of crime.** Beliefs, expectations, and values that minimize the harm of criminal victimization, increase the reward of crime, and reduce compliance to rules, police, and courts. Examples of attitudes supportive of crime, or procriminal attitudes, include beliefs that the police are fundamentally corrupt and that nobody gets ahead without cheating.

**criminogenic needs.** Potentially changeable characteristics of people that increase their likelihood of engaging in criminal behavior. Examples of criminogenic needs include procriminal attitudes, negative peer associations, and unemployment. See **dynamic risk factors**.

**domains.** The broad categories—psychological, interpersonal, and lifestyle—that describe the features of people and their environments that increase or decrease their likelihood of criminal behavior.

**dynamic risk factors.** Factors that contribute to risk but can change over time (e.g., social networks, thinking patterns, housing, substance use, finances, etc.), also called criminogenic needs. Dynamic factors not only add to the predictive ability of an assessment instrument, they represent those areas that can be changed through programming and interventions.

**life-course development.** The predictable pattern of human development from childhood, through adolescence, adulthood, and advanced age. The likelihood of criminal behavior is highest in adolescence and young adulthood and steadily

declines with age. People who are prone to social disruption and rule violation often show problematic behavior at multiple stages of the life course, although the nature of the problem changes (e.g., truancy during childhood, criminal convictions in youth, lifestyle instability in adulthood).

**lifestyle instability.** An inconsistent and/or chaotic pattern of daily living characterized by infrequent or nonexistent employment, high levels of substance use, unstable residence, short-term relationships, shifting priorities, and unrealistic goals.

**non-criminogenic needs.** Life problems that are worthy of intervention but are not directly related to the likelihood of criminal behavior. Examples of non-criminogenic needs include depression, sleep disorders, and poor physical health.

**r value.** In risk and needs assessment, the Pearson's  $r$  value is the measure of correlation between the risk score and recidivism. Pearson's  $r$  ranges from -1 to 1, with positive numbers indicating a positive relationship (i.e., higher risk and needs assessment scores are correlated with a higher likelihood of reoffending).

**static risk factors.** Risk factors that are unchanging or that cannot be changed through deliberate intervention (e.g., age, prior offenses). Static factors contrast with dynamic risk factors (or criminogenic needs), which can be used to inform the targets of supervision and human service interventions.



---

## Notes

1. S. L. Desmarais and J. P. Singh, *Instruments for Assessing Recidivism Risk: A Review of Validation Studies Conducted in the U. S.* (New York: The Council of State Governments Justice Center, 2013).
2. D. A. Andrews and J. Bonta, *The Psychology of Criminal Conduct, 5th ed.* (New York: Routledge, 2010); R. K. Hanson, G. Bourgon, L. Helmus, and S. Hodgson, "The Principles of Effective Correctional Treatment Also Apply to Sexual Offenders: A Meta-Analysis," *Criminal Justice and Behavior* 36 (2009): 865–91, doi: 10.1177/0093854809338545.
3. K. M. Babchishin and R. K. Hanson, "Improving Our Talk: Moving Beyond the 'Low', 'Moderate', and 'High' Typology of Risk Communication," *Crime Scene* 16, no. 1 (2009): 11–14; R. K. Hanson, K. M. Babchishin, L. Helmus, D. Thornton, and A. Phenix, "Communicating the Results of Criterion Referenced Prediction Measures: Risk Categories for the Static-99R and Static-2002R Sexual Offender Risk Assessment Tools," *Psychological Assessment* (September 12, 2016), doi: 10.1037/pas0000371.
4. M. Coligado and R. K. Hanson, "Measuring Recidivism Risk: A Survey of Practices in Canadian Corrections." Presentation, Third North American Correctional and Criminal Justice Psychology Conference, Ottawa, ON, June 2015.
5. H. E. Barbaree, C. M. Langton, and E. J. Peacock, "Different Actuarial Risk Measures Produce Different Risk Rankings for Sexual Offenders," *Sexual Abuse: A Journal of Research and Treatment* 18 (2006): 423–40, doi: 10.1177/107906320601800408.
6. J. P. Singh, S. Fazel, R. Gueorguieva, and A. Buchanan, "Rates of Violence in Patients Classified as High Risk by Structured Risk Assessment Instruments," *The British Journal of Psychiatry* 204, no. 3 (2014): 180–87, doi: 10.1192/bjp.bp.113.131938.
7. Public Safety Canada is a department of the government of Canada that was created in 2003 to ensure coordination across all federal departments and agencies responsible for national security and the safety of Canadians.
8. D. A. Andrews, J. Bonta, and R. D. Hoge, "Classification for Effective Rehabilitation: Rediscovering Psychology," *Criminal Justice and Behavior* 17 (1990): 19–52, doi: 10.1177/0093854890017001004; G. Bourgon and J. Bonta, "Reconsidering the Responsibility Principle: A Way to Move Forward," *Federal Probation* 78 (2014): 3–10.
9. Babchishin and Hanson, "Improving Our Talk: Moving Beyond the 'Low', 'Moderate', and 'High' Typology of Risk Communication"; Hanson et al., "Communicating the Results of Criterion Referenced Prediction Measures: Risk Categories for the Static-99R and Static-2002R Sexual Offender Risk Assessment Tools."
10. For statistical and empirical reasons, researchers often use odds ratios as the metric of relative risk, with odds defined as  $p/(1-p)$ , where  $p$  is a probability. For a further discussion of relative risk indicators, see R. K. Hanson, K. M. Babchishin, L. Helmus, and D. Thornton, "Quantifying the Relative Risk of Sex Offenders: Risk Ratios for Static-99R," *Sexual Abuse: A Journal of Research and Treatment* 25, no. 5 (2013): 482–515, doi: 10.1177/1079063212469060.
11. The Council of State Governments Justice Center, "A Common Language for Risk Assessment: Experts Convene in Washington," last modified September 2, 2014, <http://csgjusticecenter.org/reentry/posts/a-common-language-for-risk-assessments-experts-convene-in-washington/>.
12. R. K. Hanson and G. Bourgon, "Advancing Sexual Offender Risk Assessment: Standardized Risk Levels Based on Psychologically Meaningful Offender Characteristics" in *Risk and Need Assessment: Theory and Practice* (New York: Routledge, forthcoming).
13. C. Uggen, J. Manza, and M. Thompson, "Citizenship, Democracy, and the Civic Reintegration of Criminal Offenders," *The Annals of the American Academy of Political and Social Science* 605, no. 1 (2006): 281–310, doi: 10.1177/0002716206286898.
14. F.T. Cullen, C.L. Lonson, and D.S. Nagin, "Prisons Do Not Reduce Recidivism: The High Cost of Ignoring Science," *The Prison Journal* 9 (2011): 48s–65s, doi: 10.1177/0032885511415224; P. Gendreau, F. T. Cullen, and C. Goggin, *The Effects of Prison Sentences on Recidivism* (Ottawa, ON: Solicitor General Canada, 1999); P. Smith, C. Goggin, and P. Gendreau, *The Effects of Prison Sentences and Intermediate Sanctions on Recidivism: General Effects and Individual Differences* (Ottawa, Ontario: Public Works and Government Services Canada, 2002).

- 
15. A. Blumstein and K. Nakamura, "Redemption in the Presence of Widespread Criminal Background Checks," *Criminology* 47, no. 2 (2009): 327–357, doi: 10.1111/j.1745-9125.2009.00155.x; S. D. Bushway, P. Nieuwbeerta, and A. Blokland, "The Predictive Value of Criminal Background Checks: Do Age and Criminal History Affect Time to Redemption?" *Criminology* 49, no. 1 (2011): 27–60, 10.1111/j.1745-9125.2010.00217.x.
  16. Uggen et al., "Citizenship, Democracy, and the Civic Reintegration of Criminal Offenders."
  17. M. C. Kurlychek, S. D. Bushway, and R. Brame, "Long-Term Crime Desistance and Recidivism Patterns—Evidence from the Essex County Convicted Felon Study," *Criminology* 50, no. 1 (2012): 71-103.
  18. M. R. Durose, A. D. Cooper, and H. N. Snyder, *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010* (Washington, DC: Bureau of Justice Statistics, 2014).
  19. Cullen et al., "Prisons Do Not Reduce Recidivism: The High Cost of Ignoring Science"; Gendreau et al., *The Effects of Prison Sentences on Recidivism*; Smith et al., *The Effects of Prison Sentences and Intermediate Sanctions on Recidivism: General Effects and Individual Differences*.
  20. Bushway et al., "The Predictive Value of Criminal Background Checks: Do Age and Criminal History Affect Time to Redemption?"
  21. Durose et al., *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*; J. Bonta, T. Rugee, and M. Dauvergne, *The Reconviction Rate of Federal Offenders* (Gatineau, QC: Public Works and Government Services Canada, 2003), <http://www.publicsafety.gc.ca/cnt/rsrscs/pblctns/rcnvcn-rt-fdrl/index-eng.aspx>; S. Fazel and A. Wolf, "A Systematic Review of Criminal Recidivism Rates Worldwide: Current Difficulties and Recommendations for Best Practice," *PLoS ONE* 10, no. 6 (2015): e0130390. doi: 10.1371/journal.pone.0130390.
  22. Andrews and Bonta, *The Psychology of Criminal Conduct*.
  23. G. Bourgon and B. Armstrong, "Transferring the Principles of Effective Treatment into a 'Real World' Prison Setting," *Criminal Justice and Behavior* 32 (2005): 3–25; K. G. Sperber, E. J. Latessa, and M. D. Makarios, "Examining the Interaction between Level of Risk and Dosage of Treatment," *Criminal Justice and Behavior* 40 (2013): 338–48, doi: 10.1177/0093854812467942; M. D. Makarios, K. G. Sperber, and E. J. Latessa, "Treatment Dosage and the Risk Principle: A Refinement and Extension," *Journal of Offender Rehabilitation* 53, no. 5 (2014): 334-50, doi:10.1080/10509674.2014.922157
  24. Andrews and Bonta, *The Psychology of Criminal Conduct*.
  25. Blumstein and Nakamura, "Redemption in the Presence of Widespread Criminal Background Checks"; Bushway et al., "The Predictive Value of Criminal Background Checks: Do Age and Criminal History Affect Time to Redemption?"
  26. Bourgon and Armstrong, "Transferring the Principles of Effective Treatment into a 'Real World' Prison Setting."
  27. Bushway et al., "The Predictive Value of Criminal Background Checks: Do Age and Criminal History Affect Time to Redemption?"; Kurlychek et al., "Long-Term Crime Desistance and Recidivism Patterns—Evidence from the Essex County Convicted Felon Study"; R. K. Hanson, A. J. R. Harris, L. Helmus, and D. Thornton, "High-Risk Sex Offenders May not Be High Risk Forever," *Journal of Interpersonal Violence* 29, no. 15 (2014): 2792–813, doi: 10.1177/0886260514526062.
  28. S. D. Bushway, "Life-Course-Persistent Offenders," in *The Oxford Handbook of Criminological Theory*, eds. F. T. Cullen and P. Wilcox (New York: Oxford University Press, 2012).
  29. Bushway et al., "The Predictive Value of Criminal Background Checks: Do Age and Criminal History Affect Time to Redemption?"; Hanson et al., "High-Risk Sex Offenders May not Be High Risk Forever."
  30. Durose et al., *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*.
  31. Babchishin and Hanson, "Improving Our Talk: Moving Beyond the 'Low', 'Moderate', and 'High' Typology of Risk Communication."
  32. Durose et al., *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*.

- 
33. Y. Vergouwe, E. W. Steyerberg, M. J. Eijkemans, and J. D. Habbema, "Substantial Effective Sample Sizes Were Required for External Validation Studies of Predictive Logistic Regression Models," *Journal of Clinical Epidemiology* 58, no. 5 (2005): 475–83, doi: <http://dx.doi.org/10.1016/j.jclinepi.2004.06.017>.
  34. Andrews and Bonta, *The Psychology of Criminal Conduct*.
  35. Hanson et al., "Communicating the Results of Criterion Referenced Prediction Measures: Risk Categories for the Static-99R and Static-2002R Sexual Offender Risk Assessment Tools"; Hanson, R. K., and G. Bourgon. "Advancing Sexual Offender Risk Assessment: Standardized Risk Levels Based on Psychologically Meaningful Offender Characteristics," in *Risk and Need Assessment: Theory and Practice*, ed. Faye Taxman (New York: Routledge, forthcoming).

---

This project was supported by Grant No. 2010-MU-BX-K084 from the Bureau of Justice Assistance, Office of Justice Programs, U.S. Department of Justice, as part of the National Reentry Resource Center.

Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice, members of The Council of State Governments, Public Safety Canada, or the project's advisory group.



---

The Council of State Governments (CSG) Justice Center is a national nonprofit organization that serves policymakers at the local, state, and federal levels from all branches of government. The CSG Justice Center provides practical, nonpartisan advice and evidence-based, consensus-driven strategies to increase public safety and strengthen communities. For more about the CSG Justice Center, see [www.csgjusticecenter.org](http://www.csgjusticecenter.org).



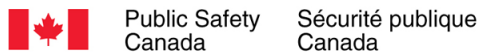
---

The Bureau of Justice Assistance (BJA), Office of Justice Programs, U.S. Department of Justice, supports law enforcement, courts, corrections, treatment, victim services, technology, and prevention initiatives that strengthen the nation's criminal justice system. BJA provides leadership, services, and funding to America's communities by emphasizing local control; building relationships in the field; developing collaborations and partnerships; promoting capacity building through planning; streamlining the administration of grants; increasing training and technical assistance; creating accountability of projects; encouraging innovation; and ultimately communicating the value of justice efforts to decision makers at every level. Visit [www.bja.gov](http://www.bja.gov) for more information.



---

The National Reentry Resource Center (NRRRC) was established in 2008 by the Second Chance Act (Public Law 110-199) and is administered by the U.S. Department of Justice's Bureau of Justice Assistance. The NRRRC provides education, training, and technical assistance to state and local governments, tribal organizations, territories, community-based service providers, non-profit organizations, and correctional institutions working to improve reentry. To learn more about the NRRRC, visit [csgjusticecenter.org/nrrc](http://csgjusticecenter.org/nrrc).



---

Public Safety Canada is a Department of the Government of Canada exercising a broad leadership role that brings coherence to the activities of the departments and agencies responsible for public safety and security. The Department's leadership role is reflected in its strategic outcome, a safe and resilient Canada, and through the pursuit of the following program activities: National Security, Emergency Management, Law Enforcement, Corrections, Crime Prevention, Border Management and Interoperability. In fulfilling its mandate, the Department works in consultation with other organizations and partners—federal departments and agencies, Provinces and Territories, non-government organizations, the private sector, foreign states, academia and communities.

Suggested citation: Hanson, R. Karl, Guy Bourgon, Robert J. McGrath, Daryl Kroner, David A. D'Amora, Shenique S. Thomas, Lahiz P. Tavarez, *A Five-Level Risk and Needs System: Maximizing Assessment Results in Corrections through the Development of a Common Language* (New York: The Council of State Governments Justice Center, 2017).

The Council of State Governments Justice Center, New York, 10007

© 2017 by The Council of State Governments Justice Center

All rights reserved.

---



ANNUAL REVIEWS **Further**

Click [here](#) to view this article's online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

# Risk Assessment in Criminal Sentencing

John Monahan<sup>1</sup> and Jennifer L. Skeem<sup>2</sup>

<sup>1</sup>University of Virginia School of Law, Charlottesville, Virginia 22903-1738; email: jmonahan@virginia.edu

<sup>2</sup>School of Social Welfare and Goldman School of Public Policy, University of California, Berkeley, California 94720-7400; email: jenskeem@berkeley.edu

Annu. Rev. Clin. Psychol. 2016. 12:489–513

First published online as a Review in Advance on December 11, 2015

The *Annual Review of Clinical Psychology* is online at [clipsy.annualreviews.org](http://clipsy.annualreviews.org)

This article's doi:  
10.1146/annurev-clipsy-021815-092945

Copyright © 2016 by Annual Reviews.  
All rights reserved

## Keywords

blameworthiness, mass incarceration, just deserts, rehabilitation, crime control, disparities

## Abstract

The past several years have seen a surge of interest in using risk assessment in criminal sentencing, both to reduce recidivism by incapacitating or treating high-risk offenders and to reduce prison populations by diverting low-risk offenders from prison. We begin by sketching jurisprudential theories of sentencing, distinguishing those that rely on risk assessment from those that preclude it. We then characterize and illustrate the varying roles that risk assessment may play in the sentencing process. We clarify questions regarding the various meanings of “risk” in sentencing and the appropriate time to assess the risk of convicted offenders. We conclude by addressing four principal problems confronting risk assessment in sentencing: conflating risk and blame, barring individual inferences based on group data, failing adequately to distinguish risk assessment from risk reduction, and ignoring whether, and if so, how, the use of risk assessment in sentencing affects racial and economic disparities in imprisonment.

## Contents

INTRODUCTION .....	490
THEORIES OF SENTENCING .....	491
Retributive Theories: “Just Deserts” .....	492
Utilitarian Theories: Crime Control .....	492
Hybrid Theories: “Limiting Retributivism” .....	493
THE ROLES OF RISK ASSESSMENT IN SENTENCING .....	493
Role 1: To Inform Decisions Regarding the Imprisonment of Higher-Risk Offenders .....	493
Role 2: To Inform Decisions Regarding the Supervised Release of Lower-Risk Offenders .....	494
Role 3: To Inform Decisions Designed to Reduce Offender Risk Status .....	494
CURRENT PRACTICE OF RISK ASSESSMENT IN AMERICAN SENTENCING .....	494
Risk Assessment in State Sentencing .....	495
Risk Assessment in Federal Sentencing .....	496
Controversies in American Sentencing .....	496
QUESTIONS OF TERMINOLOGY AND GOALS .....	497
Risk, Promotive, and Proxy Factors .....	497
Purpose, Structure, and Validation of Instruments .....	499
Selecting an Instrument .....	500
QUESTIONS OF TIMING .....	501
FOUR PERSISTENT PROBLEMS OF RISK ASSESSMENT IN SENTENCING .....	501
Problem 1: Conflating Risk and Blame Is a Category Error .....	501
Problem 2: The Virtual Impossibility of Making Individual Inferences from Group Data Is a Canard .....	505
Problem 3: Reducing Risk Is More Difficult than Assessing Risk Among Adult Offenders .....	506
Problem 4: Disparities Potentially Associated with the Use of Risk Assessment in Sentencing Are a Significant Concern .....	507
CONCLUSION .....	508

## INTRODUCTION

Since shortly after the Civil War, many American states have relied on some inchoate notion of risk assessment in criminal sentencing. New York adopted a parole statute in 1876, and Massachusetts enacted probation into law in 1878, both to be applied to offenders believed unlikely to return to crime. The explicit assessment of an offender’s risk soon became a central component of criminal sanctioning in numerous American jurisdictions. In California, for example, indeterminate sanctioning—whereby an offender was given a short minimum sentence and a long maximum one, and released from prison whenever he or she was assessed as presenting an acceptably low risk of recidivism—was introduced in 1917. In the mid-1970s, however, indeterminate sanctioning based on forward-looking assessments of an offender’s risk of committing future crime was abolished in California and elsewhere in favor of “truth in sentencing”: fixed periods of confinement based

strictly on appraisals of an offender's moral blameworthiness for the crime of which he or she has been convicted (Monahan & Skeem 2014).

Two historical trends must be appreciated to put the role of risk assessment in sentencing in context. First, the demise of risk assessment in sentencing coincided with the rise of "mass incarceration." The growth in incarceration rates in the United States since the early 1970s has been "historically unprecedented and internationally unique" (Travis et al. 2014, p. 2). One percent of the adult American population—2.4 million people—now resides in jails or prisons (Sabol et al. 2009). Western European democracies have an incarceration rate one-seventh that of the United States (Int. Cent. Prison Stud. 2013). The human and fiscal toll associated with what some have called the carceral state (Simon 2007) has become unsustainable (Cullen et al. 2011).

Second, the crime rate in the United States has plummeted since the early 1990s. According to the FBI (2014), the number of violent crimes committed per 100,000 people was 758 in 1991 and 368 in 2013—a decrease of 51%. In some cities, the crime decline is nothing short of astonishing. In New York City, for example, the homicide rate is now 18% of what it was in 1990 (Zimring 2012).

Some have suggested that these two historical trends are strongly causally related—i.e., that the rise in the rate of imprisonment produced the fall in the rate of crime. However, the National Research Council recently concluded: "The increase in incarceration may have caused a decrease in crime, but the magnitude of the reduction is highly uncertain and the results of most studies suggest it was unlikely to have been large" (Travis et al. 2014, p. 4).

Across the political spectrum (Arnold & Arnold 2015), advocates have proposed that one way to begin unwinding mass incarceration without simultaneously jeopardizing the historically low American crime rate is to put risk assessment back in sentencing. It has recently been estimated that courts in at least 20 states have begun to incorporate risk assessment "in some or all cases" of criminal sentencing (Starr 2014, p. 809). Clinical psychologists and other mental health professionals are sometimes involved in conducting clinical assessments of risk to inform sentencing decisions (Heilbrun et al. 2009). Both clinical and nonclinical psychologists are increasingly being asked to develop and validate actuarial risk assessment instruments for use by sentencing courts or parole boards.

We begin this review by sketching the major jurisprudential theories of sentencing, distinguishing those that rely on risk assessment from those that preclude it. We then characterize the varying roles that risk assessment may play in the sentencing process, and we illustrate these roles by reference to the sentencing policies of several illustrative states and proposed sentencing policies in the federal system. We clarify questions regarding the various meanings of "risk" in sentencing and questions regarding the appropriate time to assess the risk of convicted offenders. We conclude by addressing what we see as the four principal problems confronting the use of risk assessment in criminal sentencing.

## **THEORIES OF SENTENCING**

Theoretical justifications for criminal sentencing in the United States in the early decades of the twenty-first century have been aptly described by Michael Tonry (2013, p. 141) as a "crazy quilt, making it impossible to generalize about prevailing normative ideas or an 'American system of sentencing.'" Nonetheless, almost all scholars of sentencing distinguish between two broad and polar opposite approaches to the allocation of criminal punishment. One of these approaches is usually termed "retributive" or "deontological." The adherents of this approach believe that an offender's blameworthiness or culpability for crime committed in the past should be the only consideration in determining his or her punishment. The other approach is typically referred to as "consequentialist" or "utilitarian." The adherents of this approach take the position that the effect

of punishment on preventing future crime by an offender or by others should be the only concern in setting his or her punishment. Many scholars endorse some form of hybrid approach to sentencing that includes elements of both the retributive/deontological and the consequentialist/utilitarian theory. This hybrid approach is most often called “limiting retributivism” (Morris 1974). We briefly consider each approach in turn.

### Retributive Theories: “Just Deserts”

Under retributive—sometimes called deontological—theories of sentencing, as Richard Frase (2013, p. 8) has stated, “A punishment is justified according to its inherent value—whether it is a good or a bad thing in itself, regardless of whether the punishment yields good or bad consequences. Deontological principles are based on values of justice and fairness that are viewed as ends in themselves.” In the best-known retributive or deontological theory, called “just deserts,” offenders should be punished “because they deserve it, and the severity of their punishment should be proportional to their degree of blameworthiness” (Frase 2013, p. 8). Blameworthiness, in turn, consists of two components: the seriousness of the harm caused by the crime of which the offender has been convicted, and the offender’s state of mind—i.e., intent, motive, mental capacity—at the time that he or she committed it.

Whether blameworthiness for past crime is to be assessed using empirical (i.e., survey) methods or by more subjective means is a topic of active debate among retributive theorists (Robinson 2013; cf. Slobogin & Brinkley-Rubinstein 2013). Psychologists and psychiatrists sometimes play a role in sentencing systems based on retributive principles, but that role is confined to determining whether the offender’s perceived blameworthiness for crime already committed should be mitigated (e.g., due to mental illness or intellectual disability) (Melton et al. 2007). Assessing the risk of future crime plays no role in sentencing decisions based solely on backward-looking perceptions of blameworthiness.

### Utilitarian Theories: Crime Control

Under utilitarian—sometimes called consequentialist—theories, punishment is justified by recourse to its ability to decrease future criminal acts by the offender or by other would-be offenders. As Frase (2013, pp. 7–8) has elaborated:

Criminal penalties have the potential to achieve. . . crime-control effects through several mechanisms: *rehabilitation* of offenders, to address the causes of their offending; *incapacitation* of higher-risk offenders, usually by means of secure custody; specific and general *deterrence* of this and other would-be offenders, by instilling fear of punishment; and *moral education*.

Risk assessment is not relevant to deterrence or to moral education. However, both risk assessment (the incapacitation of higher-risk offenders) and risk reduction (the rehabilitation of offenders) are of central importance in forward-looking consequentialist theories. Without at least some ability to validly estimate an offender’s risk of recidivism (e.g., through the use of actuarial assessment instruments) and hopefully to reduce that level of risk (e.g., through the use of evidence-based psychological interventions), there would be few positive consequences flowing from consequential theories of sentencing.



## Hybrid Theories: “Limiting Retributivism”

Many scholars have argued that any workable theory of sentencing must address both retributive and utilitarian concerns, rather than just one of them. The most influential hybrid theory of sentencing is that proposed by Norval Morris (1974) and called “limiting retributivism.” In Morris’s theory, retributive principles set upper (and sometimes lower) limits on the severity of punishment, and within this range of what he called “not undeserved” punishment, utilitarian concerns—such as the offender’s risk of recidivism—could be taken into account. Kevin Reitz (2011, p. 472) elaborates:

Here, proportionality in punishment is understood as an imprecise concept with a margin of error, not reducible to a specific sanction for each case. The “moral calipers” available to human beings are set wide, the theory asserts, producing a substantial range of justifiable sentences for most cases. At some upper boundary, we begin to feel that a penalty is clearly disproportionate in severity and, at a lower point, we intuit that it is clearly too lenient (Morris 1974, Frase 2002). Imagining a generous spread between the two, limiting retributivism would permit utilitarian purposes to determine sentences within the morally permissible range.

Different theories of limiting retributivism might specify a broader or a narrower range of limits set by retributive concerns. A mean period of sanctioning of five years, for example, might have a permissible range of sentencing—set by backward-looking moral considerations—of four to six years, of three to seven years, or of two to eight years—ranges within which forward-looking risk assessments might be used to choose a specific sentence length. For example, Christopher Slobogin (2011, p. 1130) has articulated an extremely utilitarian model of sentencing that would have a broad range of permissible sentences, “cabined only very loosely by desert.”

It bears emphasis that the use of risk assessment under any form of this hybrid, limiting retributivism theory implies that even a very high estimated risk of future crime does not justify a sentence that exceeds the upper bound of severity perceived as morally proportionate to the crime of which the offender has been convicted. Simply put, risk assessment should not be used to sentence offenders to more time than they morally deserve.

The highly influential Model Penal Code (Tentative Draft No. 3; Am. Law Inst. 2014) explicitly adopts the hybrid, limiting retributivism approach to criminal sentencing. According to the Code, sentencing must take place “within a range of severity proportionate to the gravity of offenses, [and] the blameworthiness of offenders.” Within this range, a specific sentence must be chosen in a manner that promotes “offender rehabilitation [and] incapacitation of dangerous offenders” [§1.02(2), p. 2]. This hybrid model of sentencing is the one that we adopt here to structure our discussion of risk assessment.

## THE ROLES OF RISK ASSESSMENT IN SENTENCING

Within the constraints described above lie three important roles for risk assessment in sentencing.

### Role 1: To Inform Decisions Regarding the Imprisonment of Higher-Risk Offenders

Risk assessment can provide an empirical estimate of whether an offender has a sufficiently high likelihood of again committing crime to justify incapacitation. That is, within a range of severity set by moral concerns about the criminal act of which the offender has been convicted, risk

assessment can assist in determining whether, on utilitarian crime-control grounds, an offender should be sentenced to the upper-bound of that range (Skeem & Monahan 2011).

### **Role 2: To Inform Decisions Regarding the Supervised Release of Lower-Risk Offenders**

Risk assessment can provide an empirical estimate of whether an offender has a sufficiently low likelihood of committing additional crime to justify an abbreviated period of incapacitation, supervised release (probation/parole), or no incapacitation at all. Within a range of severity set by moral concerns about the criminal act of which the offender has been convicted, risk assessment can assist in determining whether, on utilitarian crime-control grounds, an offender should be sentenced to the lower-bound of that range (Monahan & Skeem 2014).

### **Role 3: To Inform Decisions Designed to Reduce Offender Risk Status**

Risk assessment can also inform correctional strategies to reduce an offender's risk status. Any valid tool can be used to identify higher-risk offenders to prioritize for more intensive services, placing others at appropriately lower levels of service. Programs that match the intensity of correctional services to offenders' risk level have been shown to reduce recidivism (Lowenkamp et al. 2006).

As we discuss below, some tools—in addition to estimating an offender's risk status, or likelihood of recidivism compared to other offenders—can also be used to estimate an offender's risk state, or current likelihood of recidivism compared to his or her past likelihood (Skeem & Mulvey 2002). These tools include variable risk factors that can be used to monitor ebbs and flows in an offender's risk state and adjust levels of supervision and services accordingly. As risk state increases, services and surveillance can be intensified to manage risk.

These tools also attempt to identify causal risk factors that can be changed by a given rehabilitation program and, when changed, will result in a lowering of the likelihood that the offender will commit additional crime. To the extent that causal risk factors can be identified and modified, risk assessment can do more than passively estimate or monitor an offender's likelihood of recidivism. It can actively reduce that likelihood (Dvoskin et al. 2011).

Each of these three roles for risk assessment in sentencing, if successfully accomplished, can advance the crime control objectives of the criminal law.

## **CURRENT PRACTICE OF RISK ASSESSMENT IN AMERICAN SENTENCING**

Crime control objectives have taken center stage in the current criminal justice reform movement: “From appalling incarceration numbers, budgetary crises, and greater public knowledge, momentum for reform has redirected the discussion on crime away from the question of how best to punish to how best to achieve long-term public safety” (Subramanian et al. 2014, p. 2). Over recent years, 27 states have enacted large-scale, data-based justice reinvestment efforts to use resources more efficiently and effectively by “expanding eligibility for community corrections and improving supervision, employing the use of diversion and treatment, revising sentence lengths and prioritizing prison resources” (Lawrence 2013, p. 3). Risk assessment plays an essential role in many of these state efforts—and figures prominently in proposals for sentencing reform.

In fact, the Model Penal Code (Tentative Draft No. 3; Am. Law Inst. 2014) directs sentencing commissions to develop valid actuarial instruments to estimate offenders' relative risk and treatment needs (§ 6B.09), and encourages the use of these instruments to inform decisions about

## THE PENNSYLVANIA COMMISSION ON SENTENCING

The most carefully documented work on risk assessment in sentencing has been done by the Pennsylvania Commission on Sentencing. The Commission developed an initial risk scale for select offenders (i.e., those convicted of offenses of medium severity) that consisted of eight risk factors. The factors (with scoring in parentheses) were: (1) gender (female = 0; male = 1); (2) age (less than 24 years = 3; 24–29 = 2; 30–49 = 1; 50+ = 0); (3) county (rural counties = 0; smaller urban counties = 1; Allegheny and Philadelphia counties = 2); (4) total number of prior arrests (0 arrests = 0; 1 = 1; 2 to 4 = 2; 5 to 12 = 3; 13+ = 4); (5) prior property arrests (no = 0; yes = 1); (6) prior drug arrests (no = 0; yes = 1); (7) current property offender (no = 0; yes = 1); (8) offense gravity score (4+ = 0; 1 to 3 = 1; note that *more* serious offenses, such as aggravated assault, are scored 0, and *less* serious offenses, such as writing bad checks, are scored 1).

The Commission validated the risk scale on two samples of offenders (combined  $N = 44,377$ ). In these samples, 12% of offenders scored in the “low risk” range (i.e., total scores = 0–4) and 88% did not (i.e., total scores = 5–14). Recidivism was defined as re-arrest for any crime within three years of release. Of offenders designated low risk, 22% recidivated; in comparison, 56% of non-low-risk offenders recidivated.

In June 2015, the Commission decided to exclude “county” as a factor on the risk scale. The Commission is now developing nine separate risk assessment scales for offenders with differing degrees of offense severity. The incorporation of risk assessment in criminal sentencing in Pennsylvania is still pending. Reports are available at <http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment/>.

whether to impose a community or prison sentence, particularly for “otherwise prison-bound offenders who may be safely diverted from incarceration” (p. 33).

### Risk Assessment in State Sentencing

Risk assessment has become a staple of discourse about evidence-based sentencing and corrections (Casey et al. 2011, Desmarais et al. 2015, Elek et al. 2015, Natl. Conf. State Legis. 2015) (see sidebar, The Pennsylvania Commission on Sentencing). Across the United States, statutes and regulations require that risk assessments inform individualized decisions about the appropriate level of security/supervision or services/programs for state probationers, prisoners, and parolees (Role 3 above), or mandate that risk assessments be included in parole eligibility reports or in presentence investigation reports (Roles 1–3 above).

As explained later, the most controversial applications involve front-end sentences that judges impose. A handful of states have incorporated risk assessment into sentencing guidelines as one factor that judges *may* consider in determining the appropriate sentence within the limits established by law. For example, the Virginia Criminal Sentencing Commission has developed, validated, and applied an actuarial risk assessment tool to reduce the state’s jail and prison population by 25% (Va. Crim. Sentencing Comm. 2014). A tool that distills simple risk factors (e.g., age, felony record, offense type, not regularly employed, male) is used to assess nonviolent offenders bound for incarceration under the state’s sentencing guidelines. Those who represent a low risk of reoffending are recommended for alternative punishment such as probation, jail (rather than prison), or restitution payments; offenders with higher scores proceed with their sentence recommendations unchanged. In 2014, judges sentenced 38% of low-risk offenders to an alternative punishment.

The front-end approach adopted by the Utah Sentencing Commission (2014a) focuses more explicitly on risk reduction than risk assessment. The Commission specified that “a validated risk and criminogenic needs assessment” should be conducted on all felony convictions prior to

sentencing to accurately “diagnose” the offender’s risk and needs to tailor supervision and treatment orders that can reduce recidivism. The risk-needs tool applied as part of the presentencing investigation is the 54-item Levels of Service Inventory-Revised (LSI-R) (Andrews & Bonta 1995; subscales include criminal history, antisocial attitudes/orientation, education/employment problems, and substance abuse). When imposing a sentence, the judge is encouraged to consider both the sentence calculated under the sentencing guidelines and the LSI-R-influenced recommendation of Adult Probation and Parole (Utah Sentencing Comm. 2014b).

Some states have applied risk assessment in novel ways to scaffold justice reinvestment efforts. For example, at the front end, nonviolent felony drug offenders who obtain moderate-high scores on the LSI-R and high scores on tests for drug problems are diverted from prison into community-based drug treatment programs (Kans. Sentencing Comm. 2015). At the back end, risk assessment is grafted onto efforts to shorten sentences by creating or expanding earned time credits, which allow certain inmates to accelerate their release date by participating in educational, vocational, treatment, or other risk-reduction programs (Larkin 2014, Lawrence 2009). For example, in Washington (State Wash. Dep. Correct. 2015), certain inmates may reduce their prison time by up to 50% by participating in available programs outlined in their individual reentry program, which is informed by risk and needs assessment. Earned time reductions are limited (to 10–33%) for some inmates with violent conviction offenses or relatively high risk scores.

### **Risk Assessment in Federal Sentencing**

Over recent years, multiple bipartisan bills have been introduced in Congress to reform federal sentencing—so far, to no avail. Still, pressure is building behind efforts to unwind federal mass incarceration. Of bills before Congress, the Sentencing Reform and Corrections Act of 2015 (SRCA) is the most comprehensive. This bill stitches together reforms modeled after successful state justice reinvestment efforts, including narrowing the range of offenders to whom mandatory minimum sentences apply (a front-end fix) and expanding recidivism reduction programs and early-release incentives across offenders (a back-end fix).

Risk assessment plays a role at the back end by structuring risk-reduction efforts and earned time credit. The SRCA directs the Attorney General to develop and validate a postsentencing assessment of inmates’ risks and needs, ensure that staff can reliably administer the assessment, and partner with agencies to make relevant risk-reduction programming available to inmates (from substance abuse treatment to faith-based classes). Prison staff would assess each prisoner upon admission to develop a case plan for risk reduction or—for low-risk offenders—for productive activity (e.g., prison jobs). Staff would periodically review the inmate’s progress. Inmates who successfully comply with their case plan would earn up to a 33% reduction in their prison term. In addition, low- and moderate-risk offenders would be eligible for having up to 10% of their prison term spent in home confinement. Although ineligible for earning time credit, inmates convicted of homicide, terrorism, or sex offenses would earn other incentives (e.g., commissary, visitation). Risk assessment would play no role in front-end sentencing.

### **Controversies in American Sentencing**

Former Attorney General Eric Holder (2014) has expressed hesitation about using risk assessment to inform front-end sentencing decisions, especially those involving imprisonment:

By basing sentencing decisions on static factors and immutable characteristics—like the defendant’s education level, socioeconomic background, or neighborhood—[risk assessments] may exacerbate

unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society. Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant's history of criminal conduct. They should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place.

Both Holder and the Department of Justice (Wroblewski 2014) urged the US Sentencing Commission to study whether front-end risk assessment has a disparate and adverse effect on racial minorities and the poor.

With this pointed exception, Holder (2014) celebrated the momentum building behind data-driven justice reform and specifically supported the use of risk assessment in back-end applications designed to reduce risk: "Data can help design paths for federal inmates to lower these risk assessments, and earn their way towards a reduced sentence, based on participation in programs that research shows can dramatically improve the odds of successful reentry." In Holder's view, everyone—even high-risk inmates—should have the chance to reduce his or her prison time.

The SCRA is remarkably consistent with these views. As Larkin (2014) noted, earned time statutes "have never been as controversial as sentencing laws" (p. 28). By the same token, risk assessment is less controversial when applied to scaffold risk-reduction efforts (Role 3 above) than to inform decisions about imprisonment or release (Roles 1 and 2).

## QUESTIONS OF TERMINOLOGY AND GOALS

The many and varied applications of risk assessment to sentencing are accompanied by a bewildering array of predictive factors, assessment instruments, and labels for both—risk/needs, criminogenic, static/dynamic, promotive/protective, proxy, actuarial... the list goes on. As Kraemer (2003) observed in a related context, "The absence of precise language is perhaps the major problem in current risk research" (p. 41). When research is translated to practice, the problem is amplified. Thus, we define more precisely what is meant by such basic concepts as risk and needs.

### Risk, Promotive, and Proxy Factors

**Risk factors.** A risk factor is a variable that precedes and increases the likelihood of criminal behavior (Kraemer et al. 1997). Monahan & Skeem (2014) differentiated among the four different types of risk factors for recidivism shown in **Table 1**. A fixed marker is a risk factor that cannot be changed (e.g., early onset of antisocial behavior). In contrast, both variable markers and variable risk factors can be shown to change over time. Change can be rapid (e.g., substance abuse can change daily), or slow (e.g., criminal behavior and antisocial traits change over years). Variable markers (such as age) cannot be changed through intervention, unlike variable risk factors (such as employment problems). Causal risk factors are variable risk factors that, when changed through intervention, can be shown to change the risk of recidivism.

All four types of risk factors are relevant to risk assessment (Roles 1 and 2 above). Variable markers and variable risk factors are relevant to monitoring changes in risk over time (for demonstrations, see Cohen & VanBenschoten 2014, Greiner et al. 2015, Howard & Dixon 2013, Jones et al. 2010). But only causal risk factors are directly relevant to risk reduction (Role 3 above). Put simply, treatment-relevant risk factors are causal risk factors. Unless a variable risk factor has been shown to be causal, there is little reason to assume that reducing the risk factor will reduce violence.

**Table 1 Four types of risk factors**

Type of risk factor	Definition	Example
Fixed marker	Unchangeable	Gender
Variable marker	Unchangeable by intervention	Age
Variable risk factor	Changeable by intervention	Employment status
Causal risk factor	Changeable by intervention; when changed, reduces recidivism	Substance abuse

Adapted from Kraemer et al. (1997) and Monahan & Skeem (2014).

This fact is rarely recognized in current discourse. Instead, variable risk factors have been confused with causal risk factors under the rubric of “needs,” “criminogenic needs,” or “dynamic risk factors.” The latter phrases are often misused as synonyms for causal risk factors—they typically reference risk factors that theoretically can be changed through intervention to reduce risk, but empirically have not been shown to do so. Most “needs” are variable risk factors, given the current state of evidence.

The most compelling form of evidence that a risk factor was causal would be a randomized controlled trial in which a targeted intervention was shown to be effective in changing one or more variable risk factors, and the resulting changes were shown to reduce the likelihood of posttreatment recidivism (for rare demonstrations, see Kroner & Yessine 2013). It is nearly impossible to locate such randomized controlled tests. Most correctional programs are aimed at multiple factors at the same time in a “blunderbuss fashion” (Kraemer et al. 2001, p. 854) that thwarts efforts to identify causal risk factors. Substance abuse and criminal thinking patterns have been targeted most precisely in treatment research and come closest to qualifying as causal (see Monahan & Skeem 2014).

**Promotive factors.** The principles outlined above for defining risk factors (which predict the unwelcome outcome of reoffending) also apply to promotive factors (which predict the welcome outcome of desistance from offending; see Offord & Kraemer 2000). So a promotive factor precedes and increases the likelihood of desistance and may be fixed, variable, or causal.

Promotive factors often are confused with protective factors. Promotive factors simply act in the opposite direction of risk factors (i.e., predict desistance via a main effect, across high- and low-risk cases), whereas protective factors moderate the impact of risk factors (i.e., predict desistance via an interaction, particularly in high-risk cases; Masten 2014). That is, promotive factors reduce the probability of reoffending, whereas protective factors reduce the probability of reoffending among persons exposed to risk factors (Farrington et al. 2012).

To scaffold positive risk-reduction approaches (focused on strengths rather than deficits), promotive factors have been added to some risk-assessment tools. The value of doing so is not clear. On one hand, when a promotive factor (e.g., gainful employment) is merely the polar opposite of a risk factor (e.g., unemployment), two terms are applied to the same variable, and nothing substantive is gained (Farrington et al. 2012). On the other hand, a few promising factors have emerged from the desistance literature (e.g., supportive intimate relationships, hope and self-efficacy, prosocial identity; see Serin et al. 2010, Ullrich & Coid 2011), and promotive scales have been shown to add predictive utility to risk scales and to moderate risk (Jones et al. 2015). On balance, much more (and better) research is needed before variables that robustly meet the criteria for promotive factors—much less, causal promotive factors relevant to risk reduction—can be identified.

**Proxy factors.** In the context of sentencing, risk factors are often labeled “proxies” for other variables. Frase (2014) and Harcourt (2015) have called criminal history a proxy for risk. This use of the term seems appropriate. When sentencing commissions provide a utilitarian rationale for embedding criminal history in their guidelines (e.g., criminal history identifies high-risk offenders who need to be incapacitated), they telegraph their intent to use criminal history (an indirect indicator) to approximate or stand in for risk (which is not measured directly).

Other uses seem inappropriate. Opponents of risk assessment in sentencing assert that criminal history has become a proxy for race (Harcourt 2015)—as have a number of other risk factors (e.g., employment status, education, and neighborhood are proxies for race and poverty; Starr 2014). Here, the label “proxy” conveys little more than an observation that these predictors of recidivism overlap. Criminal history and race are correlated, but it is not clear that criminal history is intended to proxy for race (i.e., to camouflage discrimination).

In our view, little light will be shed on the relation between risk factors—particularly controversial ones—unless terms such as “proxy” are operationally defined. Kraemer et al. (2001) clarify how risk factors can work together to predict an outcome such as recidivism. In their system, a proxy is a correlate of a strong risk factor that also appears to be a risk factor for the same outcome—but the only connection between the correlate and the outcome is the strong risk factor correlated with both. By their criteria, criminal history is a proxy for race only if race dominates in predicting recidivism (i.e., maximum potency in predicting recidivism is achieved by race alone—not criminal history alone, or the combination of criminal history and race). This is not the case. Because criminal history predicts recidivism more strongly than race, it will probably dominate race (Berk 2009, Bonta et al. 1998, Durose et al. 2014). Criminal history is not a proxy for race; instead, it overlaps race and possibly mediates race’s relation to recidivism.

## Purpose, Structure, and Validation of Instruments

As risk assessment has become part of mainstream corrections and sentencing, an active industry has grown up around it. Commercial off-the-shelf tools—sometimes customized to sites—have proliferated alongside government instruments designed for specific applications. This dizzying array of risk assessment tools may be ordered along three orthogonal dimensions: purpose, degree of structure, and quality of validation.

**Purpose.** Risk assessment instruments differ in the sentencing goal(s) they are meant to fulfill: Some are designed exclusively to predict recidivism (assess risk to fulfill Roles 1 and 2 above), whereas others are meant to inform risk reduction (assess needs to fulfill Role 3 above). Prediction-oriented tools (such as Virginia’s risk assessment) are designed for efficient prediction, whereas reduction-oriented tools (such as the LSI-R used in Utah) include variable risk factors to address in supervision and treatment. As the emphasis on risk reduction increases, so should the emphasis on variable (and ostensibly causal) risk factors.

In our view, distinctions between risk and needs (and associated generations of tools) create more confusion than understanding. Basically, tools differ in the sentencing goal they are meant to fulfill and in their emphasis on variable risk factors.

**Structure.** Risk assessment tools also differ in the extent to which they structure or replace professional judgment with actuarial rules and formulae (Skeem & Monahan 2011). Specifically, tools vary in whether they specify rules for generating two, three, or all four of the following components of the risk assessment process: (a) identifying empirically valid (and legally acceptable) risk factors, (b) determining a method for measuring (scoring) these risk factors, (c) establishing a

procedure for combining scores on the risk factors, and (*d*) producing an estimate of recidivism risk.

Some tools structure only the identification and measurement processes, leaving professionals to rely on their own judgment to combine scores and estimate whether an offender is low, medium, or high risk [see the Historical Clinical Risk Management-20 (HCR-20) tool; Guy et al. 2015]. Others, like the LSI-R, structure the identification, measurement, and combination of risk factors, but permit a professional override of the calculated risk estimate to recognize that rare factors outside the estimate can influence the likelihood of recidivism in a particular case. Completely actuarial tools, like the Virginia risk assessment (Farrar-Owens 2013), structure all four components of the process (see also Rice et al. 2013). Once an individual's risk has been calculated, the risk assessment process is complete.

**Validation.** Instruments used at sentencing also differ with respect to their evidence base (Desmarais et al. 2015). Although some have been rigorously studied and evaluated by independent parties, many have not. As observed by Gottfredson & Moriarty (2006), fundamental requirements for developing, cross-validating, and applying risk assessment tools are “routinely ignored and/or violated” (p. 178). These requirements are vital. Unless a tool is validated in a local system—and then periodically revalidated—there is little assurance that it works. Insufficiently trained and monitored staff may not reliably score a risk assessment tool. Variables that predict recidivism in a jurisdiction with ample services for offenders may not predict recidivism in a resource-poor jurisdiction. Similarly, when a variable becomes relatively common in the general population and loses its specificity to offending (e.g., having a tattoo, coming from a single-parent household), its utility for predicting recidivism may erode.

### Selecting an Instrument

Despite heated debate about the superiority of tools that differ in their purpose and/or structure, there is no compelling evidence that one validated tool forecasts recidivism better than another. In a meta-analysis of 28 studies that controlled well for methodological variation, Yang and colleagues (2010) found that the predictive efficiencies of nine risk assessment instruments were essentially interchangeable (see also Campbell et al. 2009). Point estimates of each instrument's accuracy tended to fall within a narrow band bounded by overlapping confidence intervals: The area under the curve (AUC) across instruments ranged from 0.65 to 0.71 (Yang et al. 2010), suggesting a 65% to 71% chance that a randomly selected recidivist obtained a higher score on the instrument than a randomly selected nonrecidivist. Although it is imperfect, the AUC is a measure of predictive efficiency that is widely applied in the risk assessment field because it facilitates comparison across studies that vary in base rates of recidivism. AUCs in the range typically observed for risk assessment tools (i.e., 0.65 to 0.71) may be viewed as medium effects (see Rice & Harris 2005).

Two factors may help explain the similar predictive performance of well-validated instruments. First, it is possible that each instrument reaches a natural limit to predictive utility, beyond which it cannot improve. Some evidence suggests that a limiting process makes recidivism impossible to predict beyond a certain level of accuracy (Coid et al. 2011). A scale can reach this limit quickly with a few maximally predictive items, before reaching a sharp point of diminishing returns. The limit can, however, be reached via alternative routes (e.g., fixed markers versus variable risk factors).

Second, well-validated tools may manifest similar performance because they tap common factors or shared dimensions of risk, despite their varied items and formats. In an innovative demonstration, Kroner and colleagues (2005) printed the items of four well-validated instruments on strips of paper, placed the strips in a coffee can, shook the can, and then randomly selected items



to create four new tools. The “coffee can instruments” predicted recidivism as well as the original instruments did. Factor analyses suggest that the instruments tap four overlapping dimensions: criminal history, an irresponsible lifestyle, psychopathy and criminal attitudes, and substance abuse–related problems. Each of these dimensions was similarly predictive of recidivism.

In our view, the choice of tool should primarily be guided by the actual purpose of risk assessment in a specific sentencing context. Given a pool of instruments that are well validated for the groups to which an individual belongs, our view is that the choice among them for use in sentencing should be driven by:

1. The ultimate purpose of the evaluation. If the ultimate purpose is to characterize an individual’s likelihood of recidivism relative to other people, then choose the most efficient instrument available. If the ultimate purpose is to manage or reduce an individual’s risk—and there is a realistic likelihood that individualized treatment services will be provided—then value may be added by choosing an instrument that includes variable risk factors (in the hope that some of these factors are causal).
2. The principle of fairness. Choose the instrument that yields the most similar predictive accuracy across groups (to minimize predictive bias) and the lowest mean score differences between groups (to minimize disparate impact; see Problem 4 below).

## QUESTIONS OF TIMING

One question with which sentencing authorities have wrestled since the late nineteenth century has been the appropriate point in time to assess an offender’s risk of recidivism for the purpose of determining the length of his or her prison sentence. There have been two basic options. The first is to perform a risk assessment at the time an offender is being sentenced, to inform the decision as to the length of sentence that the judge will impose. As described above, this is often referred to as front-end risk assessment. The second option is to sentence an offender to a largely indeterminate period of imprisonment and to perform a risk assessment later, at the time an offender is being considered for having his or her sentence terminated by means of release or parole to a noncustodial setting. This is usually referred to as back-end risk assessment (Frase 2013, Reitz 2011).

With the rise of truth in sentencing in the mid-1970s, discretionary parole and the risk assessments that guided it suffered a significant diminishment. Many states enacted determinate sentencing schemes that abolished parole entirely (Petersilia 2011). By 2000, less than one-quarter of all offenders released from American prisons gained release by means of discretionary parole (Rhine 2012, p. 632).

The American Law Institute recommended that parole boards no longer have discretion over when prisoners should be released. Postrelease supervision of former prisoners in the community—now often called reentry programming—would still be provided by parole agencies, but sentence length would be determined by a judge at the front end of the sentencing process (Am. Law Inst. 2011).

## FOUR PERSISTENT PROBLEMS OF RISK ASSESSMENT IN SENTENCING

### Problem 1: Conflating Risk and Blame Is a Category Error

Many clinical psychologists and psychiatrists have experience in risk assessment primarily by virtue of their involvement in performing evaluations for civil commitment. The legal standard

for civil commitment in virtually all American states requires two findings: mental illness and dangerousness. In the words of one illustrative state statute, in order for a person to be civilly committed, there must be “a substantial likelihood that, as a result of mental illness, the person will, in the near future, cause serious physical harm to himself or others” (Code of Virginia 2008). To mental health professionals accustomed to performing risk assessments in the context of civil commitment, the choice of risk factors in sentencing may appear baffling. They reasonably may wonder, “Why not just choose the risk factors with the highest predictive validity?”

The reason why the choice of risk factors in civil commitment seems obvious while the choice of risk factors in sentencing is fraught is because civil commitment is governed by public health law (Levin et al. 2010), whereas sentencing is governed by criminal law. Perceptions of blame play the lead role in retributive theories of sentencing, and an important role in hybrid theories that involve limiting retributivism. The role of a health-care professional, however, is to treat a patient’s existing diabetes, cirrhosis, or lung cancer, not to blame a patient for having eaten, drank, or smoked too much. The tension between backward-looking retributivism and forward-looking utilitarianism that pervades criminal sentencing is absent from health care. In health care, utilitarian concerns about the individual patient’s prognosis are usually all that matter.

For example, the use of gender as a risk factor for recidivism in sentencing is highly contested (Starr 2014, 2015). But not to use gender as a risk factor for various health conditions would be unimaginable. Consider cancer. All cancers of the reproductive system, of course, are gender specific. No rational oncologist screening for ovarian or uterine cancer would bother screening men, nor would he or she screen women for prostate or testicular cancer. But obvious reproductive differences that constitute the nature of what is meant by sex are far from the only gender differences pertinent to health care. Gender differences in the prevalence of various diseases are more the norm than the exception. For every man diagnosed with breast cancer, 181 women are so diagnosed. For every woman diagnosed with esophageal cancer, three men receive that diagnosis (Ernberg 2012, tables 3 and 4). Whatever controversy is raised by the use of gender as a risk factor in sentencing, the failure to use gender as a risk factor in health-care decision making would be seen as flagrant malpractice.

If the choice of which risk factors to use in sentencing is not determined solely by considerations of predictive validity, as it is in health care, what other considerations come into play? In the view of many scholars of sentencing (Starr 2014, Tonry 2014), perceptions of blame not only impose an upper (and perhaps a lower) limit on permissible sentences, but also serve as an essential moral constraint on the type of risk factors that can be used to assess an offender’s likelihood of recidivism. As we argue above, the task of assigning blame for an offender’s past crime and the task of assessing an offender’s risk for future crime are orthogonal aspects of sentencing. Indeed, the limiting retributivist theory of sentencing—which attempts to take both blame and risk into account—does so only by virtue of its partitioning the decision-making process in sentencing into two autonomous components: first, the sentencer should focus on assigning blame for past crime in order to establish a range of “not-undeserved” sentences, and then the sentencer should focus on the consequences for controlling future crime by choosing a specific sentence within the established range. In this manner, the inquiries into an offender’s blame and into an offender’s risk are not so much integrated as they are sequenced.

Dealing with the orthogonal concerns of blame and risk *seriatim* is not unduly problematic when a given variable bears on both concerns to similar effect, i.e., when both concerns point in the direction of raising, or both point in the direction of lowering, the severity of a sentence otherwise given. But dealing with the orthogonal concerns of blame and risk at the same time becomes problematic when a given variable bears importantly on one of the two concerns, but is irrelevant to the other (Harcourt 2015). And dealing with the orthogonal concerns of blame and

risk in chorus becomes highly contested when a given variable bears on each of the two concerns, but to opposite effect (Morse 2015). Illustrations of each of these three possibilities follow.

**Variables that affect perceptions of blame and assessments of risk in similar ways.** The clearest example of a variable that has comparable effects on perceptions of blame and on assessments of risk is involvement in crime (Roberts & Yalincak 2014). It has long been axiomatic in the field of risk assessment that past crime is the best predictor of future crime. All actuarial risk assessment instruments reflect this empirical truism. The empirically derived California Static Risk Assessment Instrument, for example, contains 22 risk factors for criminal recidivism, fully 20 of which—all but gender and age—are indices of past crime (Turner et al. 2009).

An offender's prior involvement in crime, however, indicates not only an increased risk that the offender will commit crime in the future, it also aggravates the perception that the offender is blameworthy for the crime for which he or she is being sentenced (Roberts & von Hirsch 2010). That is, "a record of prior offenses bears *both* on the offender's deserts *and* on the likelihood of recidivism" (von Hirsch 1976, p. 87; emphases added).

The existence of prior criminal convictions is not the only risk factor that reflects an offender's involvement in crime. Committing crime while under the influence of drugs such as methamphetamine, being a member of a violent gang, or being convicted of the current crime while under legal restraint (i.e., while on probation, parole, or bail) all reflect the depth of an offender's involvement in crime (Tonry 2014) and are often used simultaneously to aggravate perceptions of blame for past crime and to increase assessed risk of future crime.

Of course, it has long been known that prior criminal convictions can reflect the differential selection of given groups by police to arrest, by prosecutors to indict, and by judges and juries to convict—and not just the differential involvement of given groups in crime (Blumstein 1993). The extent to which this is the case is highly contested (Frase 2014) in current debates on sentencing policy (see Problem 4, below).

**Variables that affect either perceptions of blame or assessments of risk, but not both.**

Demographic and life history variables that characterize an offender may have significant predictive validity in assessing his or her likelihood of recidivism, but no bearing on the ascription of blame for the crime of which he or she was convicted. Consider first demography. Both race and gender correlate significantly with criminal recidivism (Blumstein et al. 1986, Durose et al. 2014). However, neither race nor gender is seen as bearing on an offender's blameworthiness for having committed crime—as a class, offenders who are women are seen as no more (or no less) blameworthy than offenders who are men, and offenders who are African American are seen as no more (or no less) blameworthy than offenders who are white. As Frase (2014) has argued, settled law has taken one of these demographic variables off the table for use as a risk factor in sentencing:

Race is really in a class by itself. The history of de jure racial discrimination in the United States, and continuing de facto discrimination, make race a highly "suspect" criterion, especially when it is used to support policies that disfavor minorities and favor whites (which is the most likely scenario in the sentencing context). . . [R]ace can never be given any formal role in issues of sentencing severity even if it is found to be correlated with and predictive of risk. (p. 149)

The law is much less settled with respect to the use of an offender's gender as a risk factor in sentencing, however. One of us has argued that using gender as a risk factor for recidivism should have little difficulty surviving legal challenge (Monahan 2006). Starr (2014), on the other hand, recently has written that using gender as a risk factor in sentencing "raises serious constitutional

concerns, and. . . is also troubling on policy grounds” (p. 806). Lay opinion on this issue appears to cleave as sharply as academic commentary: In a recent survey, Scurich & Monahan (2015) found that approximately half the respondents were open to the possibility of using gender as a risk factor for criminal recidivism, and half were not.

The use of life history variables in sentencing has received even less legal attention than the use of demographic ones. Whether a convicted offender has completed high school or is employed are predictively valid risk factors for recidivism (Farrington & Ttofi 2011, Tanner-Smith et al. 2013) and are frequently included on risk assessment instruments used in sentencing (Pa. Comm. Sentencing 2011). But educational attainment and employment status do not bear on an offender’s blameworthiness for having committed crime. A high school dropout is no more (or no less) blameworthy than a high school (or college) graduate when he or she decides to commit a crime. The same can be said of people with or without a job. The developers of many risk assessment instruments appear to believe that it is acceptable to use an offender’s life decisions as risk factors in sentencing. Others strongly disagree. According to one influential scholar (Tonry 2014),

Free citizens are. . . entitled to decide to seek university degrees, join apprenticeship programs, or live lawfully hand-to-mouth as many artists, musicians, and writers do by some combination of choice and necessity. Citizens are entitled to choose not to work at all and to live on income from trust funds or indulgent parents. . . Many offenders, however, do not—in a fundamental sense—choose to be poorly housed, poorly employed or unemployed, and poorly educated. Some do. Even if poor peoples’ choices are more constrained than those of more privileged people, they are lawful choices all the same. (p. 174)

**Variables that affect perceptions of blame and assessments of risk in opposite ways.** The clearest example of a variable that has opposite effects on perceptions of blame and on assessments of risk is combat-induced trauma. Elbogen et al. (2014), in a large study of veterans who served in Iraq and Afghanistan, found that combat experience and resulting posttraumatic stress disorder were among the strongest risk factors for a soldier’s perpetration of serious violence to others. Combat-induced trauma, therefore, can function as a risk factor for recidivism and therefore serve to increase the severity of a criminal sentence otherwise given. According to no less an authority than the United States Supreme Court, however, such trauma can also function to mitigate the offender’s blameworthiness for the commission of crime, and therefore serve to reduce the severity of the criminal sentence otherwise given.

In *Porter v. McCollum* (2009), the Court unanimously held that the trial counsel of a defendant who was a decorated war veteran had rendered ineffective assistance by failing to present his client’s military service as a mitigating factor in sentencing. “Our Nation,” the Court stated, “has a long tradition of according leniency to veterans in recognition of their service, especially for those who fought on the front lines as Porter did. Moreover, the relevance of Porter’s extensive combat experience is not only that he served honorably under extreme hardship and gruesome conditions, but also that the jury might find mitigating the intense stress and mental and emotional toll that combat took on Porter” (*Porter v. McCollum* 2009). In response to *Porter*, the federal Sentencing Guidelines (US Sentencing Comm. 2010, § 5H1.11) were revised to permit military service to be invoked in arguing for a downward departure from the sentence recommended by the Guidelines. The Commentary to the Guidelines states that “courts have often considered the impact military service has on the individual before the court; sometimes courts impose more lenient sentences when, in the court’s view, the defendant suffers from a mental or emotional condition that is traceable to the defendant’s military service” (p. 13).

Our purpose in this section is not to call various risk factors for recidivism “in” or “out” for use in sentencing. Rather, we have attempted to sharply distinguish risk assessment in the context of sentencing from risk assessment in the more familiar public health context of civil commitment, and to describe how, in the view of many scholars, perceptions of blame morally constrain not just the range of possible sentences, but also the nature of the risk factors that can be used to sentence an offender within this range. As we have argued, the task of assigning blame for an offender’s past crime and the task of assessing an offender’s risk for future crime are orthogonal aspects of sentencing. At the end of the day, however, someone—a judge, or perhaps a parole board—must join these two concerns together to form a single value on a continuous dimension of sentencing severity. The way in which these concerns are united is a matter of great debate. To date, this debate has largely been confined to the fields of law and philosophy. We hope to widen the conversation to include psychologists and other mental health professionals.

## **Problem 2: The Virtual Impossibility of Making Individual Inferences from Group Data Is a Canard**

The issue that in recent years has generated more controversy than any other in the field of risk assessment is Hart and colleagues’ (2007) provocative thesis that the margins of error surrounding individual risk assessments of violence are so wide as to make such predictions “virtually meaningless” (p. 263). As later stated by Cooke & Michie (2010), “On the basis of empirical findings, statistical theory, and logic, it is clear that predictions of future offending cannot be achieved, with any degree of confidence, in the individual case” (p. 259) (see also Hart & Cooke 2013).

Since its first publication, the Hart et al. thesis has been vigorously contested (Harris et al. 2008). For example, Hanson & Howard (2010) state that the wide margin of error for individual risk assessments is a function of having only two possible outcomes (violent or not violent) and therefore conveys nothing about the predictive utility of a risk assessment tool. Because all violence risk assessment approaches, not just actuarial approaches, yield some estimate of the likelihood that a dichotomous outcome will occur, none are immune from Hart et al.’s argument (as Hart et al. recognize). Indeed, their thesis, “if true . . . would be a serious challenge to the applicability of any empirically based risk procedure to any individual for anything” (Hanson & Howard 2010, p. 277).

Contrary to the thesis of Hart and colleagues, our view (Faigman et al. 2014, 2015; Monahan & Skeem 2014) is that group data theoretically can be, and in many areas empirically are, highly informative when making decisions about individual cases, including decisions about sentencing. Consider two examples from risk assessment in other areas. In the insurance industry, “until an individual insured is treated as a member of a group, it is impossible to know his expected loss, because for practical purposes that concept is a statistical one based on group probabilities. Without relying on such probabilities, it would be impossible to set a price for insurance coverage at all” (Abraham 1986, p. 79). In weather forecasting, “extensive statistical data are available on the average probability of the events [meteorologists] are estimating,” and therefore when meteorologists “predict a 70% chance of rain, there is measurable precipitation just about 70% of the time” (Natl. Res. Council. 1989, p. 46).

Mossman (2015) uses a medical analogy rather than one from insurance or meteorology:

Suppose a 50-year-old man learns that half of people with his diagnosis die in five years. He would find this information very useful in deciding whether to purchase an annuity that would begin payouts only after he reached his 65th birthday. Similarly, if all one knew about an individual was his Static-99R score [Hanson, Babchishin, Helmus & Thornton 2013] and that he came from a population for which

the Static-99R data and rates were relevant, the individual's Static-99R score would be the best and the only basis for making a probabilistic judgment about his future behavior. This is true even though many factors not considered by the Static-99R (e.g., employment status, substance use, and family relationships) affect a sex offender's likelihood of recidivism. (p. 99)

The recent and meticulous critique of the Hart et al. series of articles by two world-class statisticians finally may have laid this controversy to rest. Imrey & Dawid (2015) conclude that Hart et al.'s "technical statistical arguments against actuarial risk estimation are simply fallacious." In their view, the thesis of Hart, Michie, and Cooke

misconceives the nature of actuarial risk estimation and the source of its espoused benefits. In principle, precise estimation of individual risk is not needed for ARAIs [i.e., Actuarial Risk Assessment Instruments], or any other risk assessment method, to provide great benefit. If groups of individuals with high and low propensities for violence recidivism can be distinguished, and courts act upon such distinctions, recidivism will decline to the extent that groups most prone to violence are incapacitated, and infringements upon those least so prone are minimized. And both society and offenders will be better served even if we cannot be sure, based on tight statistical intervals, from precisely which individual offenders this betterment derives. (Imrey & Dawid 2015)

### **Problem 3: Reducing Risk Is More Difficult than Assessing Risk Among Adult Offenders**

Although a wealth of empirical guidance is available for assessing adult offenders' risk of recidivism, far less is available for reducing that risk. As explained previously, risk factors known to be causal are in short supply: With the possible exception of substance abuse and criminal thinking patterns, there is no compelling evidence that changing particular risk factors reduces recidivism.

In truth, variable risk factors are the best point of reference the field has to offer for reducing risk. In a randomized controlled trial, Bonta et al. (2011) found that, compared to untrained probation officers, specially trained probation officers spent more time discussing variable risk factors with their probationers (e.g., criminal thinking patterns, antisocial associates), and their probationers were less likely to reoffend. This provides indirect support for the principle of targeting variable risk factors to reduce risk, but certainly does not specify which factors are causal. So the field is left with blunderbuss interventions aimed at a "variety of influences, some of which. . . dilute or divert from intervention effects that derive from changing causal risk factors" (Kraemer et al. 2001, p. 854).

If causal factors are in short supply, high-quality adult correctional services are rare indeed. Based on a cohort of California prisoners, Petersilia & Weisberg (2010) found that substance abuse treatment (of any sort) was offered to ten percent of those with substance abuse problems, and basic anger control treatment was offered to one-quarter of one percent of those with anger problems. Evidence-based treatment programs and principles are even more scarcely implemented in adult correctional settings (Lowenkamp et al. 2006).

Still, efforts are being made to turn the *Titanic*. As part of the evidence-based sentencing movement, agencies are taking systematic action to provide offenders with access to promising types of programming (Casey et al. 2011). For example, probation agencies are developing their own treatment resources (e.g., cognitive-behavioral groups)—and using validated checklists to assess the extent to which community treatment providers adhere to known principles of effective correctional intervention. Creating infrastructure for risk reduction will be challenging but necessary to realize any modicum of success.

## Problem 4: Disparities Potentially Associated with the Use of Risk Assessment in Sentencing Are a Significant Concern

According to the most recent data from the Bureau of Justice Statistics (Carson 2014), young (i.e., 18- to 19-year-old) black males are over nine times more likely than young white males to be imprisoned. As Frase (2013) has stated:

Even when such disparity results from the application of seemingly appropriate, race-neutral sentencing criteria, it is still seen by many citizens as evidence of societal and criminal justice unfairness; such negative perceptions undermine the legitimacy of criminal laws and institutions of justice, making citizens less likely to obey the law and cooperate with law enforcement. (p. 210)

The question here is whether the use of risk assessment in sentencing affects racial disparities in imprisonment. As noted previously, Former Attorney General Eric Holder believes that it does: Although risk assessments “were crafted with the best of intentions, I am concerned that they may inadvertently undermine our efforts to ensure individualized and equal justice” (Holder 2014).

Whether risk assessment affects sentencing disparities is an important empirical question. Risk assessment could exacerbate sentencing disparities, as Holder hypothesizes. But risk assessment could also reduce or have no effect on disparities. Given findings in the general sentencing literature (Ulmer 2012), the effect of risk assessment on disparities is probably conditioned on contextual factors. It may vary, for example, as a function of the baseline sentencing context and the instrument chosen. Consider each possibility in turn.

First, whether risk assessment exacerbates, ameliorates, or has no effect on disparities is a question anchored to the baseline sentencing context, i.e., risk assessment compared to what? Racial and socioeconomic disparities depend on where one is sentenced (Ulmer 2012), so—holding all else constant—the effect of risk assessment on disparities depends on what practices are being replaced. Although practices will vary, common denominators include (a) judges’ intuitive and informal consideration of offenders’ likelihood of recidivism, which is less transparent, consistent, and accurate than evidence-based risk assessment, and (b) sentencing guidelines that heavily rely on criminal history and have been shown to contribute heavily to racial disparities (Frase 2009).

Second, the effect of risk assessment on disparities may depend on the instrument chosen. On utilitarian grounds alone, any instrument used to inform sentencing must be shown to predict recidivism with similar accuracy across groups. That is, the instrument empirically must be free of predictive bias (in statistical terms, race must not moderate the instrument’s predictive utility). However, given a pool of instruments that are free of predictive bias, some instruments will yield greater mean score differences between groups than others (Skeem & Lowenkamp 2015). Although such instruments with greater group differences are not biased, their use at sentencing arguably will have greater disparate impact (in legal terms) or inequitable social consequences (in moral terms; Reynolds & Suzuki 2012).

In short, much more research is needed to define the conditions under which risk assessment affects sentencing disparities. Studies can determine, for example, how strongly different instruments correlate with race, which risk factors drive that correlation, and what (if anything) can be done to reduce the correlation without compromising predictive utility. Guidance is available from similar efforts undertaken in related fields (e.g., tests of differential item functioning by racial groups for cognitive tests used in education; Reynolds 2000). If policymakers blindly eradicate risk factors from a tool because they are contentious, they risk reducing predictive utility and exacerbating the racial disparities they seek to ameliorate. It may be politically tempting, for example, to focus a tool tightly on criminal history because this variable is associated with perceptions of

blameworthiness and is also easily assessed by referring to conviction records. But risk estimates based on a broader set of factors predict recidivism better than criminal history and tend to be less correlated with race (Berk 2009, Skeem & Lowenkamp 2015). Such estimates also provide more points of reference for risk reduction efforts.

## CONCLUSION

The past several years have seen a remarkable surge of interest in the use of risk assessment in criminal sentencing (Hamilton 2015a,b). Political advocates who agree on little else have coalesced in proposing that the way to unwind mass incarceration in America without jeopardizing the country's historically low crime rate is to make risk assessment much more prominent in sentencing criminal offenders. Several pioneering states have already incorporated risk assessment in sentencing for some or all convicted offenders. Many other states and the federal government are actively debating whether they, too, should implement what is increasingly being referred to as evidence-based sentencing. As these debates ensue, the questions underscored in this review will be front and center. Which predictively valid risk factors are morally and legally acceptable to include in risk assessment instruments? When should those instruments be administered to convicted offenders? How can the criminal justice system promote the reduction of risk and not merely its assessment? Will a revived emphasis on recidivism risk exacerbate, ameliorate, or leave unaffected the enormous racial and economic disparities that have long characterized the American penal system? Our hope is that psychological science can play a major advisory role at what may be a historic crossroad in American sentencing policy.

### SUMMARY POINTS

1. The need to unwind mass incarceration without jeopardizing public safety is fueling interest in the use of risk assessment to inform sentencing. Across America, states are using risk assessment to inform decisions about the imprisonment of higher-risk offenders, the supervised release of lower-risk offenders, and the treatment of offenders in efforts to reduce risk.
2. Risk assessment is relevant to utilitarian (crime control), but not retributive (just deserts), sentencing concerns. Ideally, retributive concerns set a permissible range for the sentence, and risk assessment is used to select a particular sentence within that range. Risk assessment should not be used to sentence offenders to more time than they morally deserve.
3. The retributive task of assigning blame for past crime and the utilitarian task of assessing risk for a future crime are orthogonal. It is difficult to integrate these orthogonal concerns to determine an offender's sentence when a given factor bears on only one concern (e.g., male gender increases risk but not blameworthiness) or bears on both concerns in opposite directions (e.g., combat-induced trauma both increases risk and can mitigate blameworthiness). Perceptions of blameworthiness constrain the risk factors perceived as appropriate to consider at sentencing.
4. Clear conceptualizations and precise terminology are needed to advance the use of risk assessment at sentencing. We provide operational definitions for specific types of risk, promotive, and proxy factors. Our analysis indicates that there is much more empirical direction for assessing risk than for reducing risk. Risk factors known to be causal are in short supply.



5. A significant concern is whether the use of risk assessment will exacerbate, mitigate, or have no effect on racial disparities in imprisonment. The answer to this question may vary with the baseline sentencing context (i.e., risk assessment compared to what?) and the instrument chosen (i.e., degree of predictive bias or disparate impact).
6. Group data are informative when making sentencing decisions about individual cases.
7. Although validated risk assessment instruments vary in their purpose and structure, they have similar levels of accuracy in predicting recidivism. Given a pool of instruments that are well validated for the groups to which an individual belongs, the choice among them for use in sentencing should be driven by the ultimate purpose of the evaluation (i.e., risk assessment versus risk reduction) and the principle of fairness (i.e., degree of predictive bias or disparate impact).

## ACKNOWLEDGMENTS

The authors are grateful to Kimberly Kessler Ferzan and Cynthia Kempinen for their insightful comments.

## LITERATURE CITED

- Abraham KS. 1986. *Distributing Risk: Insurance, Legal Theory, and Public Policy*. New Haven, CT: Yale Univ. Press
- Am. Law Inst. 2011. *Model Penal Code: Sentencing (Tentative Draft No. 2)*. Philadelphia: Am. Law Inst.
- Am. Law Inst. 2014. *Model Penal Code: Sentencing (Tentative Draft No. 3)*. Philadelphia: Am. Law Inst.
- Andrews DA, Bonta J. 1995. *Levels of Service Inventory—Revised*. Toronto, ON: MHS Inc.
- Arnold J, Arnold L. 2015. Fixing justice in America. *Politico Mag*. <http://www.politico.com/magazine/story/2015/03/criminal-justice-reform-coalition-for-public-safety-116057.html>
- Berk R. 2009. The role of race in forecasts of violent crime. *Race Soc. Probl.* 1:231–42
- Blumstein A. 1993. Racial disproportionality of U.S. prison populations revisited. *Univ. Colo. Law Rev.* 64:743–760
- Blumstein A, Cohen J, Roth JA, Visher CA. 1986. *Criminal Careers and “Career Criminals.”* Washington, DC: Natl. Acad. Press
- Bonta J, Bourgon G, Rugge T, Scott TL, Yessine AK, et al. 2011. An experimental demonstration of training probation officers in evidence-based community supervision. *Crim. Justice Behav.* 38:1127–48
- Bonta J, Law M, Hanson K. 1998. The prediction of criminal and violent recidivism among mentally disordered offenders: a meta-analysis. *Psychol. Bull.* 123:123–42
- Campbell MA, French S, Gendreau P. 2009. The prediction of violence in adult offenders: a meta-analytic comparison of instruments and methods of assessment. *Crim. Justice Behav.* 36:567–90
- Carson E. 2014. *Prisoners in 2013*. Washington, DC: Bur. Justice Stat.
- Casey PM, Warren RK, Elek JK. 2011. *Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group*. Williamsburg, VA: Natl. Cent. State Courts. <http://www.ncsc.org/~media/Microsites/Files/CSI/RNA%20Guide%20Final.ashx>
- Code of Virginia. 2008. Involuntary temporary detention, § 37:2–809 (B)
- Cohen TH, VanBenschoten SW. 2014. Does the risk of recidivism for supervised offenders improve over time? Examining changes in the dynamic risk characteristics for offenders under federal supervision. *Fed. Probat.* 78:41–54
- Coid JW, Yang M, Ullrich S, Zhang T, Sizmur S, et al. 2011. Most items in structured risk assessment instruments do not predict violence. *J. Forensic Psychiatry Psychol.* 22:3–21
- Cooke DJ, Michie C. 2010. Limitations of diagnostic precision and predictive utility in the individual case: a challenge for forensic practice. *Law Hum. Behav.* 34:259–64

- Cullen FT, Jonson CL, Nagin DS. 2011. Prisons do not reduce recidivism: the high cost of ignoring science. *Prison J.* 91:48–65S
- Desmarais SL, Johnson KL, Singh JP. 2015. Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychol. Serv.* In press
- Durose M, Cooper A, Snyder H. 2014. *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*. Washington, DC: Bur. Justice Stat.
- Dvoskin JA, Skeem JL, Novaco RW, Douglas KS, eds. 2011. *Using Social Science to Reduce Violent Offending*. New York: Oxford Univ. Press
- Elbogen EB, Johnson SC, Wagner HR, Sullivan C, Taft CT, Beckham JC. 2014. Violent behaviour and post-traumatic stress disorder in US Iraq and Afghanistan veterans. *Br. J. Psychiatry* 204:368–75
- Elek JK, Warren RK, Casey PM. 2015. *Using Risk and Needs Assessment Information at Sentencing: Observations from Ten Jurisdictions*. Williamsburg, VA: Natl. Cent. State Courts. <http://www.ncsc.org/~media/Microsites/Files/CSI/RNA%202015/Final%20PEW%20Report%20updated%2010-5-15.ashx>
- Ernberg I. 2012. Cancer. In *Handbook of Clinical Gender Medicine*, ed. K Schenck-Gustafsson, P DeCola, D Pfaff, D Pisetsky, pp. 238–51. Basel, Switz.: Karger
- Faigman DL, Monahan J, Slobogin C. 2014. Group to individual (G2i) inference in scientific expert testimony. *Univ. Chic. Law Rev.* 81:417–80
- Faigman DL, Slobogin C, Monahan J. 2015. Gatekeeping science: using the structure of scientific research to distinguish between admissibility and weight in expert testimony. *Northwest. Univ. Law Rev.* In press
- Farrar-Owens M. 2013. The evolution of sentencing guidelines in Virginia: an example of the importance of standardized and automated felony sentencing data. *Fed. Sentencing Rep.* 25:168–70
- Farrington DP, Loeber R, Ttofi MM. 2012. Risk and protective factors for offending. In *The Oxford Handbook of Crime Prevention*, ed. BC Welsh, DP Farrington, pp. 46–69. New York: Oxford Univ. Press
- Farrington DP, Ttofi MM. 2011. Protective and promotive factors in the development of offending. In *Antisocial Behavior and Crime: Contributions of Developmental and Evaluation Research to Prevention and Intervention*, ed. T Bliesener, A Beelmann, M Stemmler, pp. 71–88. Cambridge, MA: Hogrefe Publ.
- FBI (Fed. Bur. Investig.). 2014. *Crime in the United States 2013*. Washington, DC: FBI
- Frase RS. 2002. Limiting retributivism. In *The Future of Imprisonment*, ed. M. Tonry, pp. 83–119. New York: Oxford Univ. Press
- Frase RS. 2009. What explains persistent racial disproportionality in Minnesota’s prison and jail populations? *Crime Justice* 38:201–80
- Frase RS. 2013. *Just Sentencing: Principles and Procedures for a Workable System*. New York: Oxford Univ. Press
- Frase RS. 2014. Recurring policy issues of guidelines (and non-guidelines) sentencing: risk assessments, criminal history enhancements, and the enforcement of release conditions. *Fed. Sentencing Rep.* 26:145–57
- Gottfredson SD, Moriarty LJ. 2006. Statistical risk assessment: old problems and new applications. *Crime Delinq.* 52:178–200
- Greiner LE, Law MA, Brown SL. 2015. Using dynamic factors to predict recidivism among women: a four-wave prospective study. *Crim. Justice Behav.* 42:457–80
- Guy LS, Kusaj C, Packer IK, Douglas KS. 2015. Influence of the HCR-20, LS/CMI, and PCL-R on decisions about parole suitability among lifers. *Law Hum. Behav.* 39:232–43
- Hamilton M. 2015a. Risk-needs assessment: constitutional and ethical challenges. *Am. Crim. Law Rev.* 52:231–91
- Hamilton M. 2015b. Adventures in risk: predicting violent and sexual recidivism in sentencing law. *Ariz. State Law J.* 47:1–62
- Hanson RK, Babchishin LM, Helmus L, Thornton D. 2013. Quantifying the relative risk of sex offenders: risk ratios for Static-99R. *Sex. Abuse* 25:482–515
- Hanson RK, Howard PD. 2010. Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law Hum. Behav.* 34:275–81
- Harcourt BE. 2015. Risk as a proxy for race: the dangers of risk assessment. *Fed. Sentencing Rep.* 27:237–43
- Harris GT, Rice ME, Quinsey VL. 2008. Shall evidence-based risk assessment be abandoned? *Br. J. Psychiatry* 192:154
- Hart SD, Cooke DJ. 2013. Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behav. Sci. Law* 31:81–102

- Hart SD, Michie C, Cooke DJ. 2007. Precision of actuarial risk assessment instruments: evaluating the margins of error of group v. individual predictions of violence. *Br. J. Psychiatry* 190:s60–65
- Heilbrun K, Hart A, Green H. 2009. Risk assessment in evidence-based sentencing: context and promising uses. *Chapman J. Crim. Justice* 1:127–43
- Holder E. 2014. *Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting*. Washington, DC: US Dep. Justice. <http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th>
- Howard PD, Dixon L. 2013. Identifying change in the likelihood of violent recidivism: causal dynamic risk factors in the OASys violence predictor. *Law Hum. Behav.* 37:163–74
- Imrey PB, Dawid AP. 2015. A commentary on statistical assessment of violence recidivism risk. *Stat. Public Policy* 2:1. <http://dx.doi.org/10.1080/2330443X.2015.1029338>
- Int. Cent. Prison Stud. 2013. *World Prison Brief*. London: Int. Cent. Prison Stud. <http://www.prisonstudies.org/world-prison-brief>
- Jones NJ, Brown SL, Robinson D, Frey D. 2015. Incorporating strengths into quantitative assessments of criminal risk for adult offenders: the Service Planning Instrument. *Crim. Justice Behav.* 42:321–38
- Jones NJ, Brown SL, Zamble E. 2010. Predicting criminal recidivism in adult male offenders: researcher versus parole officer assessment of dynamic risk. *Crim. Justice Behav.* 3:860–82
- Kans. Sentencing Comm. 2015. *Justice Reinvestment Initiative in Kansas*. Topeka: Kans. Sentencing Comm. [http://www.sentencing.ks.gov/docs/default-source/publications-reports-and-presentations/ksc\\_jri\\_report.pdf?sfvrsn=2](http://www.sentencing.ks.gov/docs/default-source/publications-reports-and-presentations/ksc_jri_report.pdf?sfvrsn=2)
- Kraemer HC. 2003. Current concepts of risk in psychiatric disorders. *Curr. Opin. Psychiatry* 16:421–30
- Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. 1997. Coming to terms with the terms of risk. *Arch. Gen. Psychiatry* 54:337–43
- Kraemer HC, Stice E, Kazdin A, Offord D, Kupfer D. 2001. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *Am. J. Psychiatry* 158:848–56
- Kroner DG, Mills JF, Reddon JR. 2005. A coffee can, factor analysis, and prediction of antisocial behavior: the structure of criminal risk. *Int. J. Law Psychiatry* 28:360–74
- Kroner DG, Yessine AK. 2013. Changing risk factors that impact recidivism: in search of mechanisms of change. *Law Hum. Behav.* 37:321–36
- Larkin PJ. 2014. Managing prisons by the numbers: using the good-time laws and risk-needs assessments to manage the federal prison population. *Harvard J. Law Public Policy* 1:1–29
- Lawrence A. 2009. *Cutting Corrections Costs: Earned Time Policies for State Prisoners*. Denver: Natl. Conf. State Legis. [http://www.ncsl.org/documents/cj/earned\\_time\\_report.pdf](http://www.ncsl.org/documents/cj/earned_time_report.pdf)
- Lawrence A. 2013. *Trends in Sentencing and Corrections: State Legislation*. Denver: Natl. Conf. State Legis. <http://www.ncsl.org/Documents/CJ/TrendsInSentencingAndCorrections.pdf>
- Levin B, Hennessey K, Petrila J, eds. 2010. *Mental Health Services: A Public Health Perspective*. New York: Oxford Univ. Press. 3rd ed.
- Lowenkamp CT, Latessa EJ, Holsinger AM. 2006. The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime Delinq.* 52:77–93
- Lowenkamp CT, Latessa E, Smith P. 2006. Does correctional program quality really matter? The impact of adhering to the principles of effective intervention. *Fed. Probab.* 5:201–20
- Masten AS. 2014. *Ordinary Magic: Resilience in Development*. New York: Guilford
- Melton GB, Petrila J, Poythress NG, Slobogin C. 2007. *Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers*. New York: Guilford. 3rd ed.
- Monahan J. 2006. A jurisprudence of risk assessment: forecasting harm among prisoners, predators, and patients. *Va. Law Rev.* 92:391–435
- Monahan J, Skeem JL. 2014. Risk redux: the resurgence of risk assessment in criminal sanctioning. *Fed. Sentencing Rep.* 26:158–66
- Morris N. 1974. *The Future of Imprisonment*. Chicago: Univ. Chic. Press
- Morse SJ. 2015. Genetics and criminal justice. In *The Oxford Handbook of Molecular Psychology*, ed. T Canli, pp. 409–25. New York: Oxford Univ. Press
- Mossman D. 2015. From group data to useful probabilities: the relevance of actuarial risk assessment in individual instances. *J. Am. Acad. Psychiatry Law* 43:93–102

- Natl. Conf. State Legis. 2015. *State Sentencing and Corrections Legislation*. Denver: Natl. Conf. State Legis. <http://www.ncsl.org/research/civil-and-criminal-justice/state-sentencing-and-corrections-legislation.aspx>
- Natl. Res. Counc. 1989. *Improving Risk Communication*. Washington, DC: Natl. Acad. Press
- Offord DR, Kraemer HC. 2000. Risk factors and prevention. *Evid.-Based Ment. Health* 3:70–71
- Pa. Comm. Sentencing. 2011. *Risk/Needs Assessment Project: Review of Factors Used in Risk Assessment Instruments*. <http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment/interim-report-1-review-of-factors-used-in-risk-assessment-instruments/view>
- Petersilia J. 2011. Parole and prisoner re-entry. In *The Oxford Handbook of Crime and Criminal Justice*, ed. M Tonry, pp. 925–52. New York: Oxford Univ. Press
- Petersilia J, Weisberg R. 2010. The dangers of pyrrhic victories against mass incarceration. *Daedalus* 130:124–33
- Porter v. McCollum*, 558 U.S. 30 (2009)
- Reitz KR. 2011. Sentencing. In *Crime and Public Policy*, ed. JQ Wilson, J Petersilia, pp. 467–98. New York: Oxford Univ. Press
- Reynolds CR. 2000. Methods for detecting and evaluating cultural bias in neuropsychological tests. In *Handbook of Cross-Cultural Neuropsychology*, ed. E Fletcher-Janzen, T Strickland, CR Reynolds, pp. 249–85. New York: Springer
- Reynolds CR, Suzuki LA. 2012. Bias in psychological assessment: an empirical review and recommendations. In *Handbook of Psychology*, Vol. 10. *Assessment Psychology*, ed. IB Weiner, JR Graham, JA Naglieri, pp. 82–113. New York: Wiley. 2nd ed.
- Rhine EE. 2012. The present status and future prospects of parole boards and parole supervision. In *The Oxford Handbook of Sentencing and Corrections*, ed. J Petersilia, KR Reitz, pp. 627–56. New York: Oxford Univ. Press
- Rice ME, Harris GT. 2005. Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law Hum. Behav.* 29:615–20
- Rice ME, Harris GT, Lang C. 2013. Validation of and revision to the VRAG and SORAG: the Violence Risk Appraisal Guide-Revised (VRAG-R). *Psychol. Assess.* 25:951–65
- Roberts JV, von Hirsch A, eds. 2010. *Previous Convictions at Sentencing: Theoretical and Applied Perspectives*. Portland, OR: Hart Publ.
- Roberts JV, Yalincak OH. 2014. Revisiting prior record enhancement provisions in state sentencing guidelines. *Fed. Sentencing Rep.* 26:177–90
- Robinson P. 2013. *Intuitions of Justice and the Utility of Desert*. New York: Oxford Univ. Press
- Sabol WJ, West HC, Cooper M. 2009. *Prisoners in 2008*. Washington, DC: Bur. Justice Stat.
- Scurich N, Monahan J. 2015. Evidence-based sentencing: public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law Hum. Behav.* In press
- Serin R, Lloyd C, Hanby L. 2010. Enhancing offender re-entry: an integrated model for enhancing offender re-entry. *Eur. J. Probab.* 2:53–75
- Simon J. 2007. Rise of the carceral state. *Soc. Res.* 74:471–508
- Skeem JL, Lowenkamp C. 2015. Risk, race, and recidivism: predictive bias and disparate impact. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2687339](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339)
- Skeem JL, Monahan J. 2011. Current directions in violence risk assessment. *Curr. Dir. Psychol. Sci.* 20:38–42
- Skeem JL, Mulvey EP. 2002. Assessing the risk of violence posed by mentally disordered offenders being treated in the community. In *Care of the Mentally Disordered Offender in the Community*, ed. A Buchanan, pp. 111–42. New York: Oxford
- Slobogin C. 2011. Prevention as the primary goal of sentencing: the modern case for indeterminate dispositions in criminal cases. *San Diego Law Rev.* 48:1127–72
- Slobogin C, Brinkley-Rubinstein L. 2013. Putting desert in its place. *Stanf. Law Rev.* 65:77–135
- Starr SB. 2014. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanf. Law Rev.* 66:803–72
- Starr SB. 2015. The new profiling: why punishing based on poverty and identity is unconstitutional and wrong. *Fed. Sentencing Rep.* 27:229–36

- State Wash. Dep. Correct. 2015. *Earned Release Time (Policy #350.100)*. <http://www.doc.wa.gov/policies/showFile.aspx?name=350100>
- Subramanian R, Moreno R, Broomhead S. 2014. *Recalibrating Justice: A Review of 2013 State Sentencing and Corrections Trends*. New York: Vera Inst. Justice <http://www.vera.org/sites/default/files/resources/downloads/state-sentencing-and-corrections-trends-2013-v2.pdf>
- Tanner-Smith E, Wilson S, Lipsey M. 2013. Risk factors and crime. In *The Oxford Handbook of Criminological Theory*, ed. F Cullen, P Wilcox, pp. 89–111. New York: Oxford Univ. Press
- Tonry M. 2013. Sentencing in America, 1975–2025. In *Crime and Justice in America, 1975–2025*, ed. M Tonry, pp. 141–98. Chicago: Univ. Chic. Press
- Tonry M. 2014. Legal and ethical issues in the prediction of recidivism. *Fed. Sentencing Rep.* 26:167–76
- Travis J, Western B, Redburn S. 2014. *The Growth of Incarceration in the United States: Exploring Causes and Consequences*. Washington, DC: Natl. Acad. Press
- Turner S, Hess J, Jannetta J. 2009. *Development of the California Static Risk Assessment Instrument (CSRA)*. UCI Cent. Evid.-Based Correct. Work Pap. <http://ucicorrections.seweb.uci.edu/files/2009/11/CSRA-Working-Paper.pdf>
- Ullrich S, Coid J. 2011. Protective factors for violence among released prisoners—effects over time and interactions with static risk. *J. Consult. Clin. Psychol.* 79:381–90
- Ulmer JT. 2012. Recent developments and new directions in sentencing research. *Justice Q.* 29:1–40
- US Sentencing Comm. 2010. *Sentencing Guidelines*. Washington, DC: US Sentencing Comm.
- Utah Sentencing Comm. 2014a. *2014 Annual Report*. Salt Lake City, Utah: Utah Sentencing Comm. <http://www.sentencing.utah.gov/AnnualReports/Sentencing2014.pdf>
- Utah Sentencing Comm. 2014b. *2014 Adult Sentencing and Release Guidelines*. Salt Lake City, Utah: Utah Sentencing Comm. <http://www.sentencing.utah.gov/Guidelines/Adult/2014%20Adult%20Sentencing%20and%20Release%20final.pdf>
- Va. Crim. Sentencing Comm. 2014. *2014 Annual Report*. Richmond: Va. Crim. Sentencing Comm. <http://www.vcsc.virginia.gov/2014AnnualReport.pdf>
- von Hirsch A. 1976. *Doing Justice: The Choice of Punishments*. New York: Basic Books
- Wroblewski J. 2014. *2014 US Department of Justice Criminal Division Annual Letter to US Sentencing Commission*. Washington, DC: US Dep. Justic. <http://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf>
- Yang M, Wong SC, Coid J. 2010. The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychol. Bull.* 136:740–67
- Zimring F. 2012. *The City that Became Safe: New York's Lessons for Urban Crime and Its Control*. New York: Oxford Univ. Press



# Contents

The Efficacy of Exposure Therapy for Anxiety-Related Disorders and Its Underlying Mechanisms: The Case of OCD and PTSD <i>Edna B. Foa and Carmen P. McLean</i> .....	1
History of the Concept of Addiction <i>Peter E. Nathan, Mandy Conrad, and Anne Helene Skinstad</i> .....	29
Conducting Clinical Research Using Crowdsourced Convenience Samples <i>Jesse Chandler and Danielle Shapiro</i> .....	53
Computerized Adaptive Diagnosis and Testing of Mental Health Disorders <i>Robert D. Gibbons, David J. Weiss, Ellen Frank, and David Kupfer</i> .....	83
Diagnostic Issues and Controversies in DSM-5: Return of the False Positives Problem <i>Jerome C. Wakefield</i> .....	105
The Importance of Considering Clinical Utility in the Construction of a Diagnostic Manual <i>Stephanie N. Mullins-Sweatt, Gregory J. Lengel, and Hilary L. DeShong</i> .....	133
Internet-Delivered Psychological Treatments <i>Gerhard Andersson</i> .....	157
Developmental Demands of Cognitive Behavioral Therapy for Depression in Children and Adolescents: Cognitive, Social, and Emotional Processes <i>Judy Garber, Sarah A. Frankel, and Catherine G. Herrington</i> .....	181
Gender Dysphoria in Adults <i>Kenneth J. Zucker, Anne A. Lawrence, and Baudewijntje P.C. Kreukels</i> .....	217
Mental Imagery in Depression: Phenomenology, Potential Mechanisms, and Treatment Implications <i>Emily A. Holmes, Simon E. Blackwell, Stephanie Burnett Heyes, Fritz Renner, and Filip Raes</i> .....	249

Resolving Ambiguity in Emotional Disorders: The Nature and Role of Interpretation Biases <i>Colette R. Hirsch, Frances Meeten, Charlotte Krahé, and Clare Reeder</i> .....	281
Suicide, Suicide Attempts, and Suicidal Ideation <i>E. David Klonsky, Alexis M. May, and Boaz Y. Saffer</i> .....	307
The Neurobiology of Intervention and Prevention in Early Adversity <i>Philip A. Fisher, Kate G. Beauchamp, Leslie E. Roos, Laura K. Noll, Jessica Flannery, and Brianna C. Delker</i> .....	331
Interactive and Mediational Etiologic Models of Eating Disorder Onset: Evidence from Prospective Studies <i>Eric Stice</i> .....	359
Paraphilias in the DSM-5 <i>Anthony R. Beech, Michael H. Miner, and David Thornton</i> .....	383
The Role of Craving in Substance Use Disorders: Theoretical and Methodological Issues <i>Michael A. Sayette</i> .....	407
Clashing Diagnostic Approaches: DSM-ICD Versus RDoC <i>Scott O. Lilienfeld and Michael T. Treadway</i> .....	435
Mental Health in Lesbian, Gay, Bisexual, and Transgender (LGBT) Youth <i>Stephen T. Russell and Jessica N. Fish</i> .....	465
Risk Assessment in Criminal Sentencing <i>John Monahan and Jennifer L. Skeem</i> .....	489
The Relevance of the Affordable Care Act for Improving Mental Health Care <i>David Mechanic and Mark Olfson</i> .....	515
<b>Indexes</b>	
Cumulative Index of Contributing Authors, Volumes 3–12 .....	543
Cumulative Index of Article Titles, Volumes 3–12 .....	548

## Errata

An online log of corrections to *Annual Review of Clinical Psychology* articles may be found at <http://www.annualreviews.org/errata/clinpsy>

# Offender Risk & Needs Assessment Instruments: A Primer for Courts



# Offender Risk & Needs Assessment Instruments: A Primer for Courts

Pamela M. Casey  
Jennifer K. Elek  
Roger K. Warren  
Fred Cheesman  
Matt Kleiman  
Brian Ostrom



Center for Sentencing Initiatives

This project was supported by Grant No. 2009-DG-BX-K030 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the SMART Office, and the Office for Victims of Crime. Points of view or opinions in this document are those of the authors and do not represent the official position or policies of the United States Department of Justice.



©National Center for State Courts, 2014

---

# OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS

## EXPERT PANEL MEMBERS

**Honorable Carl Ashley**  
Milwaukee County Circuit Court

**Stephen A. Bouch**  
Principal  
JS Bouch & Associates

**Sally Kreamer**  
Director, Fifth Judicial District  
Iowa Department of Corrections

**Edward J. Latessa, Ph.D.**  
Professor & Director  
School of Criminal Justice  
University of Cincinnati

**Geraldine Nagy, Ph.D.**  
Director  
Travis County Adult Probation

**Randy K. Otto, Ph.D.**  
Department of Mental Health Law & Florida Mental Health Institute

**Honorable Ron Reinstein**  
Judicial Consultant  
Superior Court of Arizona (Retired)

**Jacey Skinner**  
Director  
Utah Sentencing Commission

**Faye S. Taxman, Ph.D.**  
Professor  
Criminology, Law and Society Department  
George Mason University

**Gina M. Vincent, Ph.D.**  
Assistant Professor, Department of Psychiatry  
Co-Director, National Youth Screening & Assessment Project  
Center for Mental Health Services Research  
University of Massachusetts Medical School

---

---

## TABLE OF CONTENTS

Acknowledgements.....	iv
Introduction.....	1
What are risk and needs assessment instruments, and why use them? .....	4
What are some examples of risk and needs assessment instruments; how do they differ? .....	8
What are the qualities of good risk and needs assessment instruments? .....	13
What practices support sound implementation of risk and needs assessment instruments? .....	20
What are the practical considerations in selecting and using risk and needs assessment instruments? .....	27
Conclusion .....	31
Notes & References .....	32
Appendix: Risk and Needs Assessment Instrument Profiles .....	A-1

---

## ACKNOWLEDGEMENTS

The Primer benefited from the knowledge, counsel, and support of many individuals. Chief among these are the members of the Expert Panel who provided expertise and advice on a topic that continues to grow and evolve. Their combined knowledge of risk and needs assessment instruments and hands-on experience with criminal justice system reforms were significant assets to the effort.

We also thank the test developers who patiently sat through our interviews and provided documents and other references for our review. Their willingness to participate in the effort was crucial to understanding each instrument and ensuring our reviews are current.

We also had the opportunity to discuss the use of various instruments with numerous judges, probation officers, and other criminal justice stakeholders we have met on our travels to different jurisdictions. We are grateful for their insights about and experiences with the various tools which have informed the Primer's suggestions and recommendations.

Finally, we extend our appreciation to the many individuals from the Bureau of Justice Assistance, Office of Justice Programs, for their support of the project and their patience in seeing the project to completion.

---

## INTRODUCTION

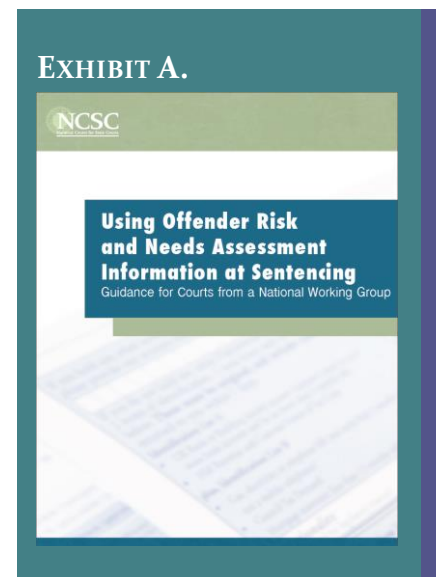
---

### *Why this Primer?*

During the last decade, the criminal justice field has focused intently on identifying programs and practices effective in reducing offender recidivism and improving public safety.<sup>1</sup> Researchers and practitioners have worked together to determine *what works best with which offenders* and, as a result, have determined that the revolving door of recidivism is not inevitable; positive outcomes for both offenders and communities are possible.<sup>2</sup> Because of the effectiveness of these evidence-based programs and practices, their use has spread to all facets of the justice system—from arrest to reentry.<sup>3</sup>

This Primer focuses on one of those decision points, *sentencing*, and on one of the tools, *risk and needs assessment (RNA) instruments*, critical to crafting sentences most likely to enhance recidivism reduction. In 2011, the National Center for State Courts published a set of guiding principles, developed by a National Working Group of practitioners and researchers, for using offender RNA information to inform sentencing decisions (Exhibit A).<sup>4</sup> The report discusses why the information is critical to the sentencing decision, how the information should be used to inform sentencing decisions, and suggestions for effectively incorporating RNA information into the sentencing process.<sup>5</sup> The guiding principles subsequently were endorsed by the Conference of Chief Justices and the Conference of State Court Administrators in a resolution, acknowledging that “research has demonstrated that the use of validated and reliable offender risk and needs assessment information to inform supervision and treatment decisions is a critical component of effective strategies to reduce recidivism.”<sup>6</sup> Specifically, the Conferences resolved to:

- “Support the National Working Group’s recommendation that offender risk and needs assessment information be available to inform judicial decisions regarding effective management and reduction of the risk of offender recidivism; and
- Endorse the guiding principles described in the National Working Group’s report as a valuable tool for state courts in crafting policies and practices to incorporate offender risk and needs assessment information in the sentencing process; and
- Encourage state and local courts to review the guiding principles and work with their justice system partners to incorporate risk and needs assessment information into the sentencing process.”



---

Although judges and other stakeholders increasingly see the value of having this information available in the sentencing process, they also have questions about how the assessments are produced and whether they are reliable, valid and fair. This Primer is a resource to help judges and others involved in sentencing understand and make knowledgeable decisions about the value and use of an assessment. It will discuss the attributes of assessment instruments that are appropriate for use in this context as well as practical considerations in selecting and properly using an RNA tool.<sup>7</sup>

The Primer also describes six of the most commonly used RNA instruments today. These descriptions are based on a review of the literature and interviews and correspondence with individuals involved in the development of the instruments. Additional research on RNA instruments is ongoing, and it is anticipated that, over time, new instruments will develop and existing instruments will be revised. Thus the criteria used for examining the six instruments in the Appendix also provide a starting point for examining any RNA instrument judges and others may consider using in their jurisdiction's sentencing process.

### *Scope of the Primer*

Practitioners use risk assessment information to inform decisions at various points in the criminal justice system. The Primer is written for judges, policy makers, and other practitioners interested in the use of RNA information at sentencing for the purpose of informing community corrections-related decisions regarding management and reduction of offender recidivism risk. It focuses on RNA instruments designed specifically to inform these community corrections-related decisions. These RNA instruments provide information relevant to sentencing considerations about an offender's amenability to supervision in the community, the level of supervision required to effectively manage the offender in the community, the types of treatment programs or other interventions most likely to reduce a specific offender's risk of reoffending, and the intensity of treatment which may be required to have recidivism-reduction effects. The Primer reviews RNA instruments that are designed for use with adult felony offenders and focused on general recidivism risk. All of the instruments provide information on an offender's risk level *and* risk factors that can be targeted with interventions to reduce recidivism.<sup>8</sup>

The Primer does not include information on instruments used exclusively at other criminal justice decision points such as pretrial release or parole, nor does it cover other instruments available to identify an offender's risk of certain types of recidivism such as violent or sexual offenses. It also does not review supplemental instruments designed to assess specialized issues such as substance abuse, mental illness, or trauma that may be warranted for use with some offenders. Some of the RNA instruments reviewed do provide additional information on offender risk at different points in the justice system (e.g., pretrial or reentry), specific types of recidivism risk (e.g., risk of committing a violent offense), or additional information regarding specific offender characteristics; but the Primer does not cover these specific aspects of the tools.<sup>9</sup> It was beyond the scope of the Primer to review all the tools focused on these aspects (e.g., all instruments

---

focused on the risk of committing a violent offense or all instruments focused on pretrial release decisions).

The remainder of the Primer covers the following five questions.

1. What are risk and needs assessment instruments, and why use them?
2. What are some examples of risk and needs assessment instruments; how do they differ?
3. What are the qualities of good risk and needs assessment instruments?
4. What practices support sound implementation of risk and needs assessment instruments?
5. What are some practical considerations in selecting and using risk and needs assessment instruments?

It is important to note that correctly using a validated RNA instrument is only one component of an evidence-based approach to reduce offender recidivism. Although the Primer is focused only on this component, readers should understand the larger context of this approach which includes, for example, matching supervision and treatment resources to an offender's risk factors, ensuring treatment programs use cognitive-behavioral skill building techniques, and selecting programs for offenders that are appropriate in light of specific offender characteristics such as gender and literacy.

---

## 1. WHAT ARE RISK AND NEEDS ASSESSMENT INSTRUMENTS, AND WHY USE THEM?

---

RNA instruments are actuarial-based tools used to classify offenders into levels of risk (e.g., low, medium, and high) and to identify and target interventions to address offender needs (e.g., antisocial attitudes, antisocial peer groups) generally related to recidivism. A RNA does not indicate whether a particular offender will actually recidivate; rather it identifies the “risk” or probability that the offender will recidivate. The probability is based on the extent to which an offender has characteristics like those of other offenders who have recidivated. For example, a RNA that results in a high risk classification means that the offender has characteristics like other offenders who have recidivated, and a low risk classification means the offender has characteristics like offenders who typically do not reoffend.<sup>10</sup>

The RNA informs risk management decisions regarding the level of supervision, i.e., the frequency and type of contact between the probation officer, client, and other individuals or agencies, required to increase the likelihood of compliance with probation conditions and ensure public safety. In addition, RNA information informs decisions regarding risk reduction strategies (e.g., cognitive behavioral programs, drug court, employment training and job assistance) that target an offender’s specific needs related to recidivism. This approach is similar to a doctor identifying a patient as a high risk for a heart attack based on several factors (e.g., high cholesterol, smoking, or poor diet) that have been shown, through research, to be related to heart disease. Although the individual may or may not actually have a heart attack, the doctor would be remiss to ignore the patient’s high risk level, and the doctor will target the patient’s treatment to those risk factors most dominant for the individual patient. Because it provides information about an offender’s relative recidivism risk and potential strategies for reducing the offender’s risk, RNA information is valuable to judges making determinations regarding an offender’s amenability to community supervision and conditions of probation in sentencing and revocation hearings.<sup>11</sup>

Research has shown the superiority of actuarial approaches to decision making over intuitive judgments in a variety of contexts, including recidivism risk.<sup>12</sup> One study of federal probation officers, for example, concluded that officers using a validated RNA tool made more consistent and accurate assessments of offender risk compared to those making unstructured professional judgments without the aid of the RNA tool.<sup>13</sup> Gottfredson and Moriarty offered several reasons for this: decision makers may not use information reliably, may not attend to base rates, may inappropriately weight predictive items, may weight items that are not predictive, and may be influenced by causal attributions or spurious correlations.<sup>14</sup> RNA instruments can assist decision makers in overcoming these issues.

To develop a RNA instrument, researchers typically collect data (or gain access to data already collected in an archive) from a representative sample of offenders on a large number of potential risk factors (e.g., criminal history, antisocial personality, school/work performance) that may be associated with recidivism. The researchers follow the offenders for a set period of time (e.g., 1-3 years) after the offenders’ prior offenses to determine whether the offenders recidivate. The data



---

from the sample of offenders are entered into a statistical model, and factors shown in the statistical model to have a significant relationship with recidivism constitute the final RNA instrument.<sup>15</sup> Subsequently, offenders who score high on the risk factors in the RNA instrument are classified as having a higher probability of reoffending; those who score lower on the risk factors are classified as having a lower probability of reoffending.

Several RNA instruments are based on the risk-need-responsivity (RNR) model. This model identifies three principles for addressing offender recidivism:<sup>16</sup>

- The **Risk** principle holds that supervision and treatment levels should match the offender's level of risk. That is, to reduce recidivism, low-risk offenders should receive less supervision and services, and higher-risk offenders should receive more intensive supervision and services.
- The **Need** principle maintains that treatment services should target an offender's dynamic risk factors or criminogenic needs (see Exhibit B<sup>17</sup>) to reduce an offender's probability of recidivism.
- The **Responsivity** principle contends that treatment interventions for offenders should use cognitive social learning strategies and be tailored to an individual offender's specific characteristics (e.g., cognitive abilities, gender) that affect successful program outcomes.

Bonta summarizes the benefit of using a RNA instrument that assists with implementing these principles:

The value of risk/need instruments is not limited to decisions around who should be supervised more closely or who should be kept in custody for the protection of the public. Because these instruments also sample criminogenic needs, they can be used to direct rehabilitation services in order to reduce offender risk.<sup>18</sup>

Research demonstrates that adherence to any one of the RNR principles correlates with a reduction in recidivism rate, and adherence to all three correlates with the highest reduction—26%—a significant decrease in current recidivism rates.<sup>19</sup> In addition to “contributing to public safety/avoiding further victimization by felony probationers and probation revocations,” the National Working Group on Using Risk and Needs Assessment Information at Sentencing highlighted several other advantages of incorporating offender assessment information into sentencing decisions:<sup>20</sup>

- Reducing prison admissions resulting from recidivism by felony probationers and probation revocations;
- Demystifying the sentencing decision and enhancing the process with scientifically-based decision tools;
- Focusing on offender accountability by requiring offenders to address their dynamic risk factors rather than placing them in programs that do not work and do not require much effort on their part;

- 
- Reducing social, economic, and family costs associated with inappropriate, and often counter-productive, interventions with low-risk offenders;
  - Ensuring sufficient prison beds for the most violent and serious offenders; and
  - Reducing prison spending by identifying offenders who can be safely and effectively supervised in the community rather than incarcerated.

#### EXHIBIT B: TERMS AND DEFINITIONS FROM THE RISK-NEEDS-RESPONSIVITY MODEL\*

Many researchers who study the link between risk factors and recidivism use RNR terms to describe various components of the link. Some of these terms are described below. Not all researchers agree with all terms and definitions. For this reason, the profile of each RNA instrument in the Appendix begins with a glossary of the terms used by the instrument's developer(s).

**Risk.** The likelihood that an offender will reoffend.

**Risk factors.** Characteristics of offenders statistically related to recidivism. Risk factors are often divided into:

- **Static risk factors.** Factors statistically related to recidivism that do not change or change in only one direction (e.g., age at first arrest, criminal history).
- **Dynamic risk factors.** Factors statistically related to recidivism that are changeable (e.g., antisocial attitudes, employment).

**Needs.** Problem areas for an offender. Needs are often divided into:

- **Criminogenic needs.** Problem areas generally related to recidivism (e.g., antisocial attitudes). These are areas typically targeted for treatment to reduce recidivism risk. Criminogenic needs and dynamic risk factors often are used interchangeably.
- **Noncriminogenic needs.** Problem areas that are not directly related to recidivism (e.g., homelessness, low self-esteem).

**Responsivity.** Targeting treatment programs to an offender's ability and learning style. Responsivity is often divided into:

- **General responsivity.** Using skill-based social learning and cognitive-behavioral programs that work to change behavior in general.
- **Specific responsivity.** Targeting treatment programs to specific offender characteristics (e.g., cognitive ability, gender).

\*Based on Andrews & Bonta (2006) and Bonta & Andrews (2007); see note 17.

Another advantage of using RNA tools is that they allow a jurisdiction to collect data over time to evaluate, for example, the effectiveness of various supervision and intervention strategies for offenders classified in different categories of recidivism risk. Data also can be used to identify the

---

types of needs most often presented by a jurisdiction's offender population and the types of supervision and intervention programs available or needed to address the needs. Thus RNA tools can also assist jurisdictions to continuously improve their allocation of resources to optimize outcome effectiveness.

---

## 2. WHAT ARE SOME EXAMPLES OF RISK AND NEEDS ASSESSMENT INSTRUMENTS; HOW DO THEY DIFFER?

---

The Appendix includes profiles of six commonly-used RNA tools:

1. Correctional Assessment and Intervention System (CAIS) which was based on the earlier Wisconsin Risk and Needs (WRN) instruments and the Client Management Classification (CMC) planning guide,
2. Correctional Offender Management Profile for Alternative Sanctions (COMPAS),
3. Level of Service Inventory-Revised (LSI-R) and Level of Service/Case Management Inventory (LS/CMI),
4. Offender Screening Tool (OST),
5. Ohio Risk Assessment System (ORAS), and
6. Static Risk and Offender Needs Guide (STRONG).

Each profile includes a glossary of terms used by the instrument developer(s) and sections on the instrument's history and current use, development, content, reliability and validity, and practical features such as automation, user qualifications, and quality assurance considerations.

The six tools include examples of instruments developed by an individual jurisdiction (i.e., OST), a state (i.e., WRN, ORAS, STRONG), or a national (i.e., CAIS, COMPAS) or provincial (LSI-R and LS/CMI) company or agency. All of the instruments have been used in multiple locations since their initial development.

RNA tools can vary in a number of ways. Several of these differences are important to an informed understanding about how a particular RNA tool may be appropriately used or implemented. Several key differences in their purpose and assessment approach follow.

### *Purpose*

As noted earlier, the Primer focuses on RNA tools developed to inform decisions about community-based supervision and treatment strategies for the general population of adult felony offenders. Several RNA tools include separate components designed for use at other decision points such as pre-trial release or release from prison (e.g., ORAS; COMPAS). Because different types of questions and outcomes are relevant for different decision points, it is important to use

### VARIATIONS IN PURPOSE AND ASSESSMENT APPROACH

- **Purpose:** How was the tool developed, for which offenders, and for which types of decisions?
- **Assessment Approach:** How does the tool calculate risk and needs; what other assessment information is provided by the tool (e.g., strengths, responsivity factors); and how is the tool administered (i.e., the methods used to conduct the assessment)?

---

any RNA tool only for the types of decision(s) for which it was intended.<sup>21</sup> Substantive differences in content may reduce predictive accuracy if a specific tool is applied at decision points other than the one at which it was originally intended for use.

RNA instruments may differ in how they define recidivism. In constructing the tools reviewed in the Primer, researchers relied on different samples of real-world offender data, outcome measures of recidivism (e.g., new arrest, conviction for a new crime, technical violation, or revocation), and follow-up periods (e.g., 1-3 years following release) for tracking reoffending.<sup>22</sup> The Community Supervision Tool (CST) of the ORAS, for example, defines recidivism as any arrest for a new crime.<sup>23</sup> The instrument developers collected data on a large number of potential RNA items from a construction sample of adult community-based offenders in Ohio and tracked new arrests over the course of a 12-month follow-up period. They retained items in the ORAS CST tool if the item correlated with rearrest during the follow-up period. The creators of the STRONG, on the other hand, examined archived, historical data on offenders released from incarceration or placed on community supervision in Washington State and defined recidivism as any subsequent felony conviction within a three year follow-up period.<sup>24</sup> Other RNA instrument developers used a more inclusive definition of recidivism, including any rule-based infraction (e.g., absconsions, rules violations, arrests, or convictions).<sup>25</sup> Differences in the type of recidivism risk calculated by a RNA tool may be meaningful in establishing local policy (or when selecting a tool to match preexisting policies), and in defining measurable recidivism reduction goals.

RNA tools may reflect the jurisdiction(s) or sample(s) of offenders on which they were developed in other ways. A RNA tool may be a valid predictor of recidivism in the particular context in which it was created, but it may not generalize well to other jurisdictions because of variations in law, policy, or the composition of the local population of adult probationers.<sup>26</sup> When one risk assessment tool originally developed in the Midwest was adopted without modification for use with probationers in New York City, researchers found that several items in the risk assessment were not related to recidivism in the New York sample.<sup>27</sup> An existing RNA tool may therefore not meet the needs of a new jurisdiction if variations in the nature or composition of the jurisdiction's target offender population alter the degree to which the instrument items and recidivism are related.<sup>28</sup>

For the above reasons, the purpose for which a RNA tool was originally designed, including the definition of recidivism used and the population on which it was developed, is an important consideration for those who use an existing RNA tool in their own jurisdiction. Subsequent validation research, if available, may also help to show that a particular RNA tool may be effectively used in a different setting or in a jurisdiction with a different demographic composition of offenders or offense types. If additional research on a particular RNA tool is not available, a good practice is to validate the instrument on the local offender population prior to adoption or full-scale implementation.<sup>29</sup>

---

## *Assessment Approach*

The researchers who created the RNA tools described in the Appendix ascribe to different theoretical approaches and different approaches to measurement. Some of the main differences among the RNA tools include (1) how they assess offender risk and needs, (2) the types of other information incorporated into the assessment, and (3) how they are administered.

**Assessment of risk and needs.** Some tools assess risk and needs together, using a single instrument and produce a composite risk and needs score, others use a single instrument and produce separate risk and needs scores, and others use separate risk and needs instruments and produce separate risk and needs scores.<sup>30</sup>

Proponents of instruments that produce a composite risk and needs score argue that all of the items in these instruments are criminogenic, i.e., they have a direct, empirically demonstrated relationship with recidivism.<sup>31</sup> In addition, because these instruments include a large proportion of items that are dynamic (i.e., changeable over time such as antisocial attitudes) as opposed to predictors that are static in nature (i.e., cannot be changed through intervention such as age), they are helpful in guiding case planning.<sup>32</sup> Assuming an instrument has been properly validated, it can help identify an offender's dynamic risk factors that, when effectively addressed, reduce recidivism risk.<sup>33</sup>

Critics of the composite score approach question the extent to which some of the dynamic risk items used in calculating the composite risk and needs score correlate with recidivism given the results of studies in different jurisdictions.<sup>34</sup> They contend that greater predictive accuracy can be achieved with shorter, more parsimonious risk scales and that separating risk and needs scales produces better measures of both.<sup>35</sup> In particular, they argue that the separate risk score is not diluted by needs items that may actually reduce the predictive ability of the risk tool. Instruments that produce a separate risk score generally rely on a smaller number (typically a dozen or less) of items found to be most predictive of recidivism in a construction sample of offenders. The separate needs score usually is based on a larger number of static and/or dynamic items that may be related to recidivism and/or identified as important by correctional officers for case management purposes.<sup>36</sup>

Critics of keeping risk and needs scores separate argue that the needs assessment portion of these RNA systems is not always subject to the same validation efforts as the risk portion.<sup>37</sup> The validated risk score is helpful in classifying an offender's risk level, but it is not helpful in identifying strategies to reduce recidivism.<sup>38</sup> Because some of the needs items may or may not be related to recidivism (e.g., items suggested by stakeholder groups as important for case planning), validation of the needs assessment is necessary to determine its effectiveness in identifying risk factors to target for intervention.

These criticisms indicate the importance for jurisdictions to look for evidence that a tool's risk and needs scores, whether provided in a composite form or separately, classify an offender correctly as low, medium, or high risk *and* also correctly identify dynamic risk factors to target for

---

risk reduction interventions.<sup>39</sup> They also provide another reason for validating any RNA instrument a jurisdiction chooses to use. Developers of RNA tools with both composite risk and needs scores, and separate risk-only scores have published research on the construction and validation of their instruments to show that each item retained in the tool has been found to be statistically related to recidivism in local construction and/or validation samples.<sup>40</sup> However, validation will ensure that the instrument retains its predictive ability when implemented in a new jurisdiction. For those instruments that provide separate needs scores, these, too, must be validated if they are to be used for identifying targets for risk reduction.<sup>41</sup>

**Other components of the assessment tool.** RNA tools also differ in the extent to which they assess other components beyond risk and needs. Some RNA tools incorporate offender strengths, also referred to as protective factors, into the assessment. A protective factor “is a variable that interacts with a risk factor to decrease the potential harmful effect of the risk factor... [acting] as a buffer that reduces the link between risk factors and later offending.”<sup>42</sup> Protective factors may include education level, employment, and the quality of family and marital relationships.<sup>43</sup> Other RNA tools include offender “responsivity factors” in the assessment. Responsivity factors are non-criminogenic offender characteristics that may affect treatment effectiveness. Responsivity factors such as the offender’s physical and mental health status, motivation to change, and learning style may affect the offender’s ability or willingness to participate in sustained treatment, or likelihood of succeeding in treatment and thus are important in case planning.<sup>44</sup>

RNA tools with separate risk and needs assessments may include both strengths (protective factors) and responsivity factors within the needs assessment. Composite RNA tools may also provide the opportunity to indicate areas of strengths (protective factors) in the full assessment (as in the LS/CMI) but separate out non-criminogenic items like responsivity factors into a different section of the tool (as in the OST).<sup>45</sup>

**Administration of the assessment.** RNA tools also differ in how they can be administered. The risk assessment component of a tool that uses separate risk and needs scales may be conducted by an intake unit using available case information and criminal records data about an offender (as with the STRONG); an interview with the offender may not be necessary.<sup>46</sup> However, the needs assessment component of such a tool and administration of composite risk-needs assessment tools both require a structured professional interview with the offender, conducted by a trained assessment administrator. Criminal records data and offender interview data may be supplemented with other methods of data collection, such as a self-report questionnaire completed by the offender undergoing assessment and/or information from collateral sources like victim statements or interviews with the offender’s family members.

Each form of assessment administration has its own pros and cons that may be weighted differently by each jurisdiction in the context of local priorities and available resources. For example, self-report surveys can be efficient, but they assume the offender understands the question being asked and also rely on the offender to supply honest answers. Structured interviews by trained professionals collect information from the offender, but in a more dynamic

---

fashion that can allow for confirmation of understanding, opportunities to probe for additional information, and a professional appraisal of the veracity of responses. Some approaches to the data review method of administration can be quite efficient: Software programs may be developed at additional cost to automate the scoring process, linking an existing data source (such as the jurisdiction's case management system) with a risk assessment application. Stakeholders should, however, be aware of the limitations of the data source upon which the risk assessment relies. For example, criminal records found in one case management system may provide only a partial picture of the offender's criminal history due to jurisdictional limitations. Stakeholders should understand the strengths and weaknesses of each data collection method and the quality of each information source(s) used by the adopted RNA tool.

To balance out shortcomings of any particular mode of assessment and as a best practice to ensure the quality of the data entering the assessment, most RNA tools require the administrator to collect information about the offender from multiple sources. For example, to obtain information about offender needs to determine appropriate treatment resources and inform case planning, a probation officer or other qualified assessment administrator will need to conduct a structured interview with the offender. Information gathered from the structured interview may be cross-checked with and/or supplemented by information provided in an offender self-report survey, a review of available records (e.g., to confirm criminal history, place of residence, educational background), and/or interviews with family members of the offender.



---

### 3. WHAT ARE THE QUALITIES OF GOOD RISK AND NEEDS ASSESSMENT INSTRUMENTS?

---

A good RNA instrument consistently produces accurate results that are fair across the types of offenders with whom the tool will be used. That is, a good tool is reliable, valid, and unbiased. Each of these general qualities is associated with specific statistical testing procedures to help ensure that the tool meets or exceeds minimum scientific standards. A description of each quality follows.

#### *Is it Reliable?*

Does the RNA tool produce consistent results if re-administered to the same person by the same or by different test administrators?<sup>47</sup> Researchers refer to this quality as reliability. Without reliability, instrument users cannot have confidence that the tool will produce an accurate result at any given time. The instrument profiles in the Appendix describe currently available research findings on the reliability of each assessment tool.

The first form of reliability referenced above – that the assessment may be administered repeatedly and produce consistent results – is called **test-retest reliability**. This form of reliability reflects the ability of the RNA instrument to generate a similar if not identical result when administered and re-administered to the same offender under the similar circumstances (i.e., by the same test administrator, assuming that nothing significant in the offender’s life has changed, for example, as a result of treatment interventions). Usually, test-retest reliability is measured using correlation statistics which show the relationship between measurements at two different points in time. Correlations range from -1.0 to +1.0, but should approach +1.0 to establish test-retest reliability. Most studies on RNA instruments do not provide information about test-retest reliability; but in broader research, scientists generally consider reliability statistics below .40 to be poor, between .40 and .59 to be fair, .60 - .74 to be good, and .75 - 1.0 to be excellent.<sup>48</sup>

The second form of reliability referenced above – that the assessment can be administered effectively by multiple test administrators – is called **inter-rater reliability** (also called inter-rater agreement). This form of reliability determines the degree to which different test administrators give the same offender similar scores on individual items as well as for the tool overall. Inter-rater reliability between two test administrators is the most common form of

#### RNA INSTRUMENT QUALITY: SIX KEY QUESTIONS

1. Is the tool reliable?
2. Is the tool valid overall?
3. Is the tool valid with all subpopulations of local offenders?
4. Is the tool easily susceptible to manipulation?
5. Has the tool been independently evaluated?
6. What are the limitations in what is empirically known about the tool?

---

reliability reported in RNA research, typically using correlation statistics.<sup>49</sup> Again, correlations may range from -1.0 to +1.0, but values should approach +1.0 to establish inter-rater reliability.

Although existing research may establish the reliability of a particular RNA tool, this information only shows that it is *possible* for the RNA tool, as it has been developed, to produce consistent results. When a tool is implemented locally, the degree of reliability *in that jurisdiction* may differ from the degree of reliability reported in prior research studies. This is because the reliability of RNA tools depends heavily on the level of training and skills of local test administrators. Both forms of reliability will be higher when test administrators receive effective, comprehensive, on-going training on how to properly use the RNA tool. Effective training will ensure that all test administrators understand the provided criteria in the same way and have the skills necessary to consistently implement established procedures when scoring the tool. Ongoing training will also help to minimize *drift* – a common tendency among test administrators to begin using the tool slightly differently from one another over time in individualistic ways that systematically distort assessment results.<sup>50</sup>

Thus when selecting and using a RNA tool, practitioners should not only be familiar with the existing research evidence demonstrating that the chosen RNA instrument is capable of achieving acceptable levels of reliability, but also understand the importance of the quality assurance mechanisms necessary to attain those levels of reliability. Those in charge of assessment should be prepared to routinely monitor reliability after the RNA tool has been implemented locally to ensure that the tool is used and continues to be used properly. This information will help determine whether the existing training package is sufficient, or if a more rigorous approach is necessary to support local use.

Reliability describes only the consistency of results generated from a RNA tool; it says nothing about how accurate those results are. Reliability is insufficient by itself to demonstrate the effectiveness of a RNA tool, but it is a necessary component of validity, which is discussed next.

### ***Is it Valid?***

The most obvious quality that a good RNA tool should have is the ability to measure what it purports to measure. This quality, called validity, focuses on measurement accuracy and also assumes that the tool can be implemented reliably (see above section).

Although validity is a singular concept, there are many different but inter-related forms of validity that reinforce one another. These multiple tests provide convergent evidence that a tool is valid. In this section, we will focus on ***predictive validity***, one of the most fundamental and important measures of validity with offender assessments.<sup>51</sup>

Predictive validity is the degree to which the results of the RNA instrument are related to behavioral outcomes of offenders *in the aggregate*. Because these testing procedures are based on averages from group data, the relationship between RNA results and behavioral outcomes for a specific individual may differ from the group results. However, group data can meaningfully

---

inform decisions about individual cases. William Grove and Paul Meehl provide the following example:

Suppose you are suffering from a distressing illness, painful or incapacitating, and your physician says that it would be a good idea to have surgeon X perform a certain radical operation in the hope of curing you. You would naturally inquire whether this operation works for this disease and how risky it is. The physician might say, "Well, it doesn't always work, but it's a pretty good operation. It does have some risk. There are people who die on the operating table, but not usually." You would ask, "Well, what percentage of times does it work? Does it work over half the time, or 90%, or what? And how many people die under the knife?"... How would you react if your physician replied, "Why are you asking me about statistics? We are talking about you – an individual patient. You are unique. Nobody is exactly like you."<sup>52</sup>

Group or aggregate data provide essential information for understanding the odds of a particular outcome. This information is applied in a number of life decisions, from more serious decisions like the medical example above to more mundane decisions like whether or not to carry an umbrella when embarking on a long walk given the local weatherman's forecast of the chance of rain. Across a number of professions and professional decision contexts, a large body of evidence demonstrates that actuarial tools produce more accurate and more reliable assessments of risk than professional judgment alone.<sup>53</sup> One of the main arguments in favor of using structured RNA tools is that, by using explicit criteria to capture information about general factors known in the scientific literature to be related to recidivism, these actuarial tools are capable of producing more consistent, accurate, objective assessments of offenders than might be generated otherwise.<sup>54</sup>

Most of the existing research on RNA instruments examines the predictive validity of the overall risk assessment component of the tool. Researchers examine the predictive validity of risk assessments empirically, using any of several different statistical techniques. The reported statistical techniques depend on the nature of the data, but at minimum will examine the relationship between the result of the assessment and a specific observed behavioral outcome (usually a form of recidivism, typically arrest or conviction for a new crime). Some of these studies also examine the extent to which each item or factor in the assessment contributes to the overall predictive validity of the risk assessment (i.e., *incremental* predictive validity).

The instrument profiles in the Appendix describe the evidence currently available on the predictive validity of each assessment tool in relation to a defined behavioral outcome (or set of outcomes). The cited evidence helps to establish the predictive validity of each tool when used under particular conditions. However, to ensure that the RNA tool is valid in a specific jurisdiction, additional local validation research is recommended. At minimum, practitioners should examine whether the tool has been validated in comparable settings with comparable target populations of offenders using the same definition of reoffending.<sup>55</sup> For a number of reasons, local validation can be helpful regardless of how often the RNA tool has been empirically validated elsewhere. Local validation research (a) will show how well the RNA tool works locally and can more concretely and convincingly demonstrate the actual benefits of using the RNA tool in that jurisdiction; (b) can help increase stakeholder confidence in the tool and encourage its

---

use; and (c) can provide invaluable research evidence to protect against potential legal challenges. Some researchers believe that local validation is required if one is seeking to adopt a RNA tool that has been validated in fewer than three similar locales.<sup>56</sup>

In the validation studies cited in the instrument profiles in the Appendix, the two most commonly reported predictive validity statistics are **correlations (*r*)** and **area-under-the-curve (AUC)** values, explained in more detail below.<sup>57</sup> Because RNA instruments classify offenders into groups of low, moderate, and high risk of recidivism to inform supervision and case planning strategies, a critical question is whether those who are classified into higher-risk groups actually show higher rates of recidivism than those classified into lower-risk groups, barring any kind of recidivism-reduction intervention. That is, an important question is not simply whether or not a risk assessment score is related to future recidivism, but whether the cutoff scores used to create the risk classification levels effectively separate low, medium, and high risk offenders.<sup>58</sup> A validation study of a good RNA tool should show the highest recidivism rates for offenders classified in the high-risk group, followed by offenders classified in the medium-risk group; the low-risk group of offenders should have the lowest recidivism rate of all.

**Correlations.** Correlations, or *r* values, are measures of association between two variables. A **point-biserial correlation ( $r_{pb}$ )** is a special kind of correlation statistic that is conducted when one of the two variables is continuous (i.e., the variable contains a range of possible values between two points, such as a risk assessment tool that generates raw scores ranging from 0 to 100), and when the other variable is dichotomous (i.e., the variable contains one of two possible values, such as when recidivism is defined as a simple yes/no to indicate whether an offender has or has not recidivated). Correlations can range from 0 to 1 (+ or -).

Correlation values provide two pieces of critical information: the direction of the relationship between two variables and the strength of that relationship. First, the sign (+ or -) indicates the direction of the relationship. In general, *r* values may be positive (“as *a* increases, *b* also increases”) or negative (“as *a* increases, *b* decreases”). All RNA tools should demonstrate an overall positive relationship with recidivism (i.e., as offender risk of recidivism scores on the RNA tool increase, actual observed recidivism should also increase). Second, the magnitude of the *r* value indicates the strength of the relationship between recidivism risk and actual recidivism. If *r* = 0, there is no relationship between recidivism risk and actual recidivism. The closer the *r* value is to 1, the stronger the relationship between the recidivism risk and actual recidivism.

Researchers will often report whether there is a “statistically significant” correlation between the raw recidivism risk scores generated by the RNA tool and offenders’ recidivistic behavior. This represents partial evidence to support a conclusion that an RNA tool does what it purports to do. However, because RNA tools are designed to produce risk level classifications, it is those classification levels – not the raw recidivism risk scores – that are actually used to inform decision-making and case planning. For this reason, better evidence of the predictive validity of a RNA tool would show that the tool accurately separates offenders into low, medium, and high risk

---

groups. A variety of statistical techniques may be used to test this, but researchers most commonly report AUC values from receiver operating characteristic analyses.

**AUC values.** AUC values represent the computed probability of the number of correct classifications, or “hits”, versus the number of incorrect classifications, or “false alarms”, by the risk assessment tool. The AUC value has advantages over other statistical techniques to help instrument users understand how well the RNA tool discriminates between offenders who will and will not reoffend, notably because it is unaffected by changes in the population’s base rate for recidivism.<sup>59</sup> An AUC = .5 means that an assessment tool is no better than chance at discriminating between recidivists and non-recidivists. The closer the AUC value is to 1, the more effective the assessment tool is at discriminating between recidivists and non-recidivists. Several groups of scientists have encouraged researchers to use and report AUCs, when possible, as the preferred measure of predictive accuracy in risk assessment, in part because the technique takes base rates into account in a standardized manner.<sup>60</sup>

When correlations and AUC values are reported as evidence for a tool’s predictive validity, researchers will interpret those values to determine how effective the tool is in practical terms. The interpretive guidelines described in Table 1 have been used by some researchers to characterize the magnitude of the “effect” of using offender risk assessment tools as small, moderate, or large.<sup>61</sup> Other researchers view these conventional guidelines as too stringent in the context of applied research and have suggested alternative cutoffs (e.g., *r* values of .1, .2, and .3 as cutoffs for small, moderate, and large effects, respectively).<sup>62</sup>

**Table 1.** General Guidelines for Interpreting Statistical Effect Sizes (Rice & Harris, 1995; 2005).

Effect	<i>r<sub>pb</sub></i>	AUC
Small	.100 to .243	.556 to .639
Moderate	.243 to .371	.639 to .714
Large	.371 or greater	.714 or greater

It is important to understand that even an effect categorized as “small” according to these conventions may meaningfully improve the assessment of risk in comparison with a business-as-usual approach.<sup>63</sup> Although scientific conventions have been established as general guidelines for interpreting the size of these effects, scientists agree that these guidelines should not be unquestioningly applied across all situations, and that “the adequacy of an assessment for a specific purpose cannot be directly inferred from single effect size indicators.”<sup>64</sup> Rather, interpreting the strength of an effect depends on a number of important factors, including but not limited to the social context of the study (e.g., what does local leadership consider to be a meaningful reduction in recidivism?) and the specific constraints of a particular research design. In fact, Rice and Harris have gone so far as to suggest that “the field of risk assessment place little reliance on plain language verbal labels because of the considerable disagreement about what they mean” among scientists, and that “clarity is best reflected by numerical characterization.”<sup>65</sup> For that reason, the Primer presents only the numerical values in the profiles of individual RNA

---

tools and does not attempt to characterize the size of the effects through a categorical label such as small, moderate, or large.

### ***Additional Validity Issues***

A good RNA tool must also produce fair results that are not systematically biased against particular subgroups of offenders and that cannot be easily manipulated by an offender to achieve desirable outcomes. Both of these concerns are subsumed within the broader research concept of validity, but merit special consideration because of the additional steps researchers must take to address these issues. Both of these issues are addressed below.

**Is it valid for all offender populations?** The instrument must produce fair and unbiased results across all of the groups of offenders on which the RNA tool will be used. This aspect of fairness is called ***differential validity***. Although the overall predictive validity of the RNA tool may have been established generally among a broad and diverse group of offenders, further examination of the predictive validity of the tool among various offender subgroups (e.g., by gender, race, ethnicity) may reveal significant differences in the degree of accuracy observed.<sup>66</sup> For example, it is possible for a risk assessment to have strong predictive validity overall, yet produce less accurate results for female offenders. Female offenders often score artificially higher (i.e., tend to be overclassified) on risk assessments that were developed with the male offender in mind and validated primarily on samples of male offenders.<sup>67</sup> Without adjustments—such as by establishing separate cutoff scores for classifying male versus female offenders as low, medium, or high risk to reoffend—tools that erroneously and systematically overclassify female offenders as higher risk will likely result in the over-supervision of female offenders in a jurisdiction that follows an evidence-based community supervision model. Moreover, some scientists have criticized the use of so-called “gender-neutral” tools with female offenders more broadly, claiming that the reliance on primarily male offender data in the instrument development process results in a tool that inadequately captures the unique criminogenic needs of female offenders.<sup>68</sup> To address these types of issues, a few providers of RNA systems now offer gender-responsive supplements in addition to the original gender-neutral version (e.g., LS/CMI, COMPAS).<sup>69</sup>

There are similar concerns regarding the predictive validity of RNA instruments for different race and ethnic groups. The extent of research on this issue varies across instruments and for different race and ethnic groups. The instrument profiles in the Appendix discuss the current research available on each tool’s predictive validity across different offender groups.

**Is it susceptible to manipulation?** Offenders may be motivated to respond artificially in ways that make them look good (called ***social desirability response bias***). Instrument developers typically incorporate strategies in the assessment process that minimize the influence of socially desirable responses on assessment results. Whether information is gathered by a trained assessment administrator conducting a structured interview with an offender or via a paper and pencil self-report measure that is completed directly by the offender, the assessment administrator is typically required to corroborate disclosed information by verifying with

---

collateral sources (e.g., official records, interviews with family or friends of the offender).<sup>70</sup> Some RNA tools with a self-report component take this a step further. The COMPAS system, for example, includes additional items in the self-report component of the assessment process that comprise what tool developers refer to as the “Lie Scale”.<sup>71</sup> These additional items are used to identify offenders who may be attempting to manipulate the results of the assessment through socially desirable responses, or what they call “faking good.” Evidence of a social desirability bias on the part of the responding offender indicates that self-reported information should be interpreted with caution and will likely require additional corroboration before RNA results can be trusted.<sup>72</sup>

### ***Additional Considerations When Reviewing Research on Risk and Needs Assessment Tools***

When reviewing the available research on a particular RNA tool, practitioners should consider two additional factors.

First, practitioners should take note of who conducted the research. Most of the available research on RNA tools has been conducted by the instrument developers themselves. Practitioners should review the research literature to determine whether the tool has been ***independently evaluated***.<sup>73</sup> That is, practitioners should determine whether the RNA tool has been rigorously evaluated by researchers who are not financially or otherwise personally invested in the success of the tool and, if so, whether those research findings support or contradict conclusions drawn by the instrument developers. Instrument developers may have an inherent conflict of interest when it comes to evaluating the success of their own tool. A bias in favor of their own tool might influence their work, consciously or not, to produce findings that cannot be reliably replicated by others. Moreover, instrument developers have more intimate knowledge about how the tool should be used that may influence how it is implemented in their testing site or how the validation study is conducted in ways that the typical user or independent researcher may not be able to duplicate from documented sources. For these reasons, it is always helpful to know whether existing research descriptions about the reliability, validity, and fairness of a tool have been replicated by others.

Second, practitioners should also understand the broader limitations of what is known about a particular tool. In researching the above psychometric properties of available RNA tools, practitioners will learn that the amount and quality of empirical research conducted varies, sometimes substantially, among the different instruments. RNA tools that have been in use longer, such as the LSI-R, will—and should—have been subjected to more rigorous evaluations and meta-analyses (analyses of the results of multiple studies) and should be supported by more documented evidence of their psychometric properties. However, simply because one RNA instrument has been studied more comprehensively than another does not necessarily mean it is a more valid tool than more recent instruments. Practical considerations, such as the resources needed to support more rigorous validation, may influence a decision about whether to use a well-studied older tool or a promising newer one. Some additional practical considerations are discussed in the next section.

---

## 4. WHAT PRACTICES SUPPORT SOUND IMPLEMENTATION OF RISK AND NEEDS ASSESSMENT INSTRUMENTS?

---

Use of a validated RNA tool is a necessary but not sufficient condition to ensure effective community-based sentencing practices. Line staff also must be equipped with the knowledge and skills necessary to use the tool properly, and management must ensure that line staff administer the tool correctly and consistently over time. A rigorous quality assurance program, including initial and ongoing staff training, coaching or mentoring, routine data monitoring, and fidelity testing (i.e., ensuring that the RNA tool is administered as it was designed), should be instituted to ensure effective implementation.

This section further discusses the importance of instrument validation and quality assurance, and key considerations at each step.

### *Instrument Validation*

**Purposes of validation.** Validation is essential to demonstrate the predictive accuracy of a RNA tool. As discussed in Section 3, the RNA tool must be supported by empirical research demonstrating that it meets basic scientific accuracy requirements in the prediction of rearrest, reconviction, or other recidivism measure of interest. Any sentencing or treatment decisions based on a RNA tool which grossly misclassifies the risk levels of offenders may not simply fail to improve outcomes; they may actually do harm to the offender. For truly high-risk offenders, less intensive supervision and treatment interventions may be ineffective.<sup>74</sup> And mandating truly low-risk offenders into more intensive supervision and/or treatment services may actually increase their recidivism risk.<sup>75</sup>

Thus a jurisdiction should not implement a RNA tool without evidentiary support that the tool appropriately categorizes the types of offenders with which the tool will be used into groups exhibiting clearly distinct probabilities of recidivism.<sup>76</sup>

Instrument validation is not only important to ensure that decision making is informed by sound data, but also to establish stakeholder confidence in the RNA tool. If probation officers, judges, and other stakeholders do not trust that the tool will enhance decision-making effectiveness, they may not use or implement the tool as intended, thereby undermining the validity of the tool. In

### PRACTICES TO SUPPORT SOUND IMPLEMENTATION

- Use a validated RNA instrument, for which validity has been demonstrated generally, is established locally, and is re-established periodically.
- Provide comprehensive initial & ongoing refresher training to all stakeholders on how to properly administer the RNA tool and understand and use its results. Develop an internal capacity to train so that these practices are sustainable.
- Routinely monitor RNA administrators for fidelity regarding proper use of overrides and consistency in scoring.



---

Maricopa County, Arizona, for example, part of the impetus for developing the OST was the observation that probation officers were not implementing the prior RNA tool as instructed because they did not believe that the tool was helpful in decision-making: Probation officers completed the tool simply because it “had to be” done.<sup>77</sup> Validation studies can provide stakeholder groups with concrete empirical evidence of the instrument’s functional value with the local offender population. This information may help to secure stakeholder buy-in when introducing evidence-based policies and practices for the first time or when integrating a new RNA instrument into existing practices. Judges and other stakeholders are more likely to support institutional changes if persuasive evidence supporting those decisions is also shared.<sup>78</sup>

**Local vs. general validation.** Instrument validity may be established locally (i.e., by commissioning a validation study within the jurisdiction in which the tool will be used) or by referencing a general body of existing validation research. A review of the existing research literature will help to determine whether or not the tool has already been validated for use in similar locations or with similar types of offenders as in one’s own jurisdiction. Some of the more established and more popular RNA tools have the benefit of a long history of research on instrument validity in an array of contexts, in a number of different jurisdictions, and conducted by a number of independent researchers. In some cases, the vendor or instrument developer warehouses data from all validation studies and can reference this data bank to determine the need for local validation. Some scientists and practitioners have indicated that if the RNA tool was developed for use with a similar population and has been validated multiple times in similar settings, or, regardless of the population on which it was developed, has been validated in at least three different jurisdictions with a similar population, setting, and definition of reoffending, local validation is not required.<sup>79</sup>

Jurisdictions can still benefit greatly from validating the chosen RNA tool locally even if instrument validity has been established generally. The same scientist-practitioner group that indicated that local validation may not be necessary in certain cases also recommends that validity still be assessed locally for any RNA tool of the type reviewed in this Primer.<sup>80</sup> As previously mentioned, differences in policy, procedure, or the makeup of the offender population may alter the predictive accuracy of a RNA tool. A local validation study will (a) inform any modifications that must be made to the content of the tool to optimize predictive validity in the local jurisdiction and ensure that it meets basic minimum scientific standards, and (b) inform the development of appropriate cutoff values for categorizing offenders into different risk levels based on actual observed differences in the probability of reoffending within the local population (also called *norming*). Judges and probation officers will be reassured that they are using a scientifically supported tool appropriate for their jurisdiction that can be confidently defended as objective, valid, and reliable. In Washington State, for example, where Department of Corrections officers may be civilly liable for their case plan decisions, a tool validated statewide offers a sense of security and protection against such liability.<sup>81</sup> This information is useful as long as the nature of the statewide sample on which the tool is validated mirrors the local population on which it is used. When properly validated, stakeholders can more confidently speak to the accuracy of the

---

classification schemes in use; the RNA tool and decisions predicated on information provided by the tool will be able to withstand critical examination.

**Revalidation.** Periodic revalidation studies of the RNA tool may also be necessary, particularly following any significant changes in local law, policing, composition of the community, or other factors that could impact offense rates or alter the common types of offending over time. Recommendations vary regarding how frequently revalidation studies should take place: One RNA instrument researcher interviewed recommended conducting revalidation studies at periodic intervals of every 3-5 years, and another instrument developer indicated that the frequency of revalidation work needed may depend on the type of assessment instrument used.<sup>82</sup> RNA tools that were developed based on emerging statistical trends observed in the relationship between existing offender data (usually convenience data like criminal history and other readily available information) and recidivism, for example, may have less stable predictive validity than RNA tools which capture information on the kinds of characteristics identified in the broader research literature as associated with criminal behavior. This is because changes in the nature and rate of recidivism on which these “statistically developed” tools are predicated, and in other factors such as contemporary community supervision practices, may reduce the predictive validity of the original assessment tool over time. In a reexamination of the original Wisconsin risk assessment tool, for example, researchers found that changes to the items and weights of the original instrument and adjustments to the risk level cutoff scores were needed in order to support continued confidence in the predictive validity of the tool.<sup>83</sup> A periodic review of classification practices will help determine whether any changes or “recalibrations” to the tool are necessary to ensure continued accuracy and appropriate classification of the local offender population over time.<sup>84</sup>

### ***Implementation Quality***

In addition to ensuring scientific support for the validity of the RNA tool, a jurisdiction should install a comprehensive plan to ensure that all users implement the tool according to its design. Without assurance of implementation quality, even a good RNA tool can produce poor or, at best, inconsistent outcomes.<sup>85</sup> A rigorous quality assurance program will not only include comprehensive and sustainable training for assessment administrators and for all users of assessment information, but also include routine quality assurance monitoring and periodic fidelity (or reliability) testing of assessment results. These components are discussed below.

### **Comprehensive and sustainable training.**

- ***Initial training and internal capacity to train.*** Users of commercially available RNA tools are generally required to undergo initial training on proper usage of the tool and the associated software before they are permitted to administer the tool. For most commercially available tools, external providers typically offer a basic two- to three-day initial training package, which covers the minimum training necessary to administer the tool. These providers also offer “train the trainers” programs to allow local jurisdictions to develop the capacity to

conduct standard trainings internally, as well as specialized courses designed to boost supplemental skills (e.g., courses on motivational interviewing, effective case planning). Establishing an internal capacity to train may be helpful not only in creating a sustainable training program for instrument administrators, but also in creating a training program to educate judges, attorneys, and other stakeholders who receive RNA information. Educating stakeholders on when and how RNA information may be appropriately applied in decision-making is a critical component of implementation that should not be overlooked, as they must understand the prescribed uses and limitations of RNA information in order to apply this knowledge effectively.<sup>86</sup> See the instrument profiles in the Appendix for details on training requirements and packages for each instrument.

If adopting or using a non-proprietary tool without an established or prepackaged training program, a training program will need to be developed from the ground up before the RNA tool can be installed. Those charged with developing the training program to support RNA installation should be knowledgeable about training strategies that optimize skill development and increase the likelihood that trained skills will be applied in practice. In a broad synthesis of implementation research literature, some researchers cited general estimates that only about 10% of trained material is typically retained by trainees.<sup>87</sup> Behavioral change is much more likely when staff members are provided with meaningful opportunities to directly apply trained skills in practical scenarios and to obtain feedback or coaching guidance for improving performance. When theory and discussion are augmented with demonstration, practice, feedback, and on-the-job coaching, 95% of trained material is retained and put into practice (see Table 2).<sup>88</sup>

**Table 2.** Summary of a Meta-Analysis of the Effects of Training and Coaching on Teachers’ Implementation in the Classroom (Joyce & Showers, 2002; excerpted from Fixsen et al., 2005, p. 30).

Training Components	Outcomes		
	(% of participants demonstrating knowledge and new skills in a training setting, and using new skills in the classroom)		
	Knowledge	Skill Demonstration	Use in Classroom
Theory and Discussion	10%	5%	0%
+ Demonstration in Training	30%	20%	0%
+ Practice & Feedback in Training	60%	60%	5%
+ Coaching in the Classroom	95%	95%	95%

- *Other ongoing training efforts.* Periodic booster or refresher training is important to prevent a problem commonly referred to as **drift**, in which test administrators start to use the same RNA tool slightly differently from one another over time in individualistic ways that distort assessment results and reduce accuracy. To prevent drift in how the RNA tool is administered and used over time, experts recommend that staff receive **refresher** (or **booster**) **training** every six months.<sup>89</sup> Refresher training should cover assessment administration as well as guidance on interpreting the results of the RNA assessment for use in supervision and case planning. Some form of refresher training is necessary not just for assessment administrators,

---

but for all those who receive RNA information for use in decision-making (i.e., probation officers in the field, judges, attorneys). As indicated above, ongoing on-the-job coaching or mentoring strategies may help to support high-fidelity implementation.<sup>90</sup> Some jurisdictions also utilize other strategies, such as peer support meetings or case review round-table meetings, to encourage users to discuss and constructively problem-solve implementation challenges.<sup>91</sup>

### **Quality assurance monitoring.**

- *Use of administrative overrides.* Some jurisdictions have specific offense-based policies in place for supervision of particular types of offenders (e.g., sex offenders) regardless of the assessed risk level of the offender, and may refer to these blanket policies as **policy overrides** of the RNA results. Typically, policy-based overrides prioritize other purposes of supervision such as risk management rather than recidivism reduction. This section focuses on overrides that occur as a result of an assessment administrator's subjective decision in an individual case based on his or her own professional judgment, or **administrative overrides**.

Most RNA tools contain a discretionary administrative override function that the assessment administrator is authorized to use to modify individual RNA results. That is, if the administrator believes that certain information about the offender is not adequately captured in the assessment and that the results should be altered to better reflect this information, the administrator may make a discretionary decision in that case to modify the offender's RNA results accordingly. Most instrument developers emphatically caution against frequent use of the administrative override function and encourage a practice in which such exceptions are made in no more than 10% of all cases (overall or per assessor).<sup>92</sup> Some instrument providers recommend a lower exception rate (e.g., 2-3%).<sup>93</sup>

To date, little research exists to document the impacts of discretionary administrative overrides on the predictive accuracy of risk assessment tools. Across a number of decision-making contexts, however, the exercise of subjective judgment by a clinician or other professional with specialized expertise in the absence of an actuarial tool or other structured decision aid, referred to as **unstructured professional judgment**, generally produces results inferior to judgments informed by these tools because humans are simply not very good at reliably and accurately identifying and weighing the complex factors that inform risk.<sup>94</sup> Frequent use of the administrative override function in an assessment tool based on the administrator's professional judgment risks diminished assessment accuracy: Studies outside of the offender risk assessment field have demonstrated that human judgment, when used *only to amend the results of an actuarial model*, still reduced predictive accuracy compared with the unmodified actuarial results.<sup>95</sup> Similarly, one recent offender risk assessment study examined the use of the professional override function in administering the LS/CMI with a sample of sex offenders.<sup>96</sup> The study showed that administrators were much more likely to apply a discretionary override to LS/CMI results in order to increase the offender's risk level than to decrease it. Importantly, the application of administrative overrides served to decrease

---

the accuracy of the assessment overall, but especially so when overriding to increase the offender's risk level. The authors discovered that administrators intuitively based their discretionary override decisions on offender characteristics that are not truly associated with recidivism risk. Although this study was conducted on a unique sample of adult felony offenders, until more research on the use of administrative overrides with the general population of felony offenders becomes available, it highlights the potential risks inherent in the practice of overriding assessment scores and serves as a caution against frequent use of the override function.

A high rate of overrides among risk assessment administrators may be indicative of more fundamental implementation problems. It may signal, for example, that more training is required on how to properly administer the RNA tool and should trigger targeted coaching, mentoring, retraining, or other quality assurance efforts. Alternatively, liberal use of the override function may be a symptom of a different problem: that staff users have low confidence in the utility of the RNA tool. In that instance, assuming the tool has been properly validated, additional efforts to educate staff on the research supporting use of the tool may be needed. (See *Instrument Validation* in this section, above.)

To deter frequent and inappropriate use of the override function, court and probation leaders have taken different approaches. Some jurisdictions permit the use of an administrative override in exceptional circumstances only, and have established protocols requiring clear documentation of reasoning and formal approval by a supervisor.<sup>97</sup> Alternatively, other jurisdictions have elected to prohibit administrative overrides entirely.<sup>98</sup>

- *Data monitoring.* A good quality assurance program should include two main efforts. First, the jurisdiction should be able to show that as a result of training, different RNA instrument administrators are able to produce consistent scores on the RNA tool and its individual items. That is, an individual should receive the same RNA results regardless of the administrator conducting the assessment. As discussed in Section 3, this type of inter-rater reliability has significant implications for the validity and credibility of the tool. A properly validated RNA tool will be supported by evidence that it *can* be scored consistently to produce reliable results. Inter-rater reliability tests will show whether the tool is being *administered* correctly by staff in the local jurisdiction, and whether the reported results from use of the validated RNA tool can be trusted.

Second, the jurisdiction should be able to identify staff members who are using the RNA tool according to established procedure and those who may require additional training or other supportive services to build the required assessment skills. Supervisors may conduct **case audits**, a periodic review of line staff assessment and scoring practices, to ensure adherence to established protocol. Supervisors may also observe and critique samples of assessment interviews in person or on audio or video tapes to provide line staff with performance feedback.<sup>99</sup> In addition, aggregate data monitoring procedures may be helpful. Some

---

researchers suggest that jurisdictions examine data collected over time to determine, for example, whether the percentage of assessed offenders who fall into each risk category (low, moderate, high) are approximately equal; whether the distribution across risk categories differs substantially between females and males or between offenders of different racial or ethnic backgrounds (which may trigger further examination of the potential for bias in application of the RNA); and whether the proportion of overrides applied exceeds maximum limits recommended by instrument developers in cases overall or in cases supervised by any individual assessment officer.<sup>100</sup>

Some RNA service providers may offer trainings or add-ons to automated RNA systems designed to support fidelity testing. For example, the ORAS includes a feature which allows the client to draw random samples of cases for internal review, and clients may complete a certification course offered by the University of Cincinnati to develop internal capacity to conduct routine fidelity studies.<sup>101</sup> If a fidelity testing software program is not available through the RNA provider, local users should be able to export data from an automated RNA system for manual analysis. If internal capacity does not exist to analyze data for quality assurance purposes, the RNA provider or other research contractors may be available to provide research services.

---

## 5. WHAT ARE THE PRACTICAL CONSIDERATIONS IN SELECTING AND USING RISK AND NEEDS ASSESSMENT INSTRUMENTS?

---

Several practical considerations will likely inform decisions about selecting and using a RNA instrument. These include considerations related to the availability of services designed to support implementation and maintenance of the RNA tool, associated costs, and the ease of use.

### *Availability of Support Options*

Some vendors may operate as a “one-stop shop,” offering not only the RNA tool itself, but also the research and training services as described in section 4 that are necessary to support quality implementation and on-going maintenance of the system over time. Vendors may conduct validation and fidelity studies, and provide train-the-trainer and user training programs to support the use of the tool. They may also establish forums for users of the tool to submit questions to instrument developers, ask questions of their peers in the community, and share information on associated policies, procedures, and practices. Vendors may also offer a range of specialized software packages that may be tailored to the needs of the client jurisdiction. The software will, at a minimum, compute the results of the assessment and generate individual assessment reports, saving time and minimizing user error. Other software options typically bundle a case management system with the automated assessment. In addition to a case planning function, these systems enable the tracking of offender outcomes and may include a variety of customizable aggregate report generation options. The case management system may be housed by the vendor on a remote server that requires local users to have internet access and assigned user login information. Often, the software bundle may be purchased and installed on a server owned and operated by the local jurisdiction. Most vendors also offer technology solutions to integrate the RNA software bundle with a client jurisdiction’s existing case management system.

#### RNA INSTRUMENTS: PRACTICAL CONSIDERATIONS

- **Availability of Support Options:** What services (e.g., RNA and reporting software, custom IT integration, user training, train-the-trainer training, quality assurance monitoring, validation research) does the RNA vendor provide? Alternatively, what support services are not available?
- **Costs:** What costs are associated with implementation and ongoing use of the RNA tool (e.g., instrument & software subscription costs, initial & ongoing stakeholder training costs, quality assurance protocol development & monitoring costs, periodic validation research)?
- **Ease of Use:** How easy is the tool to implement, administer, and use to inform decision-making?

---

Other vendors may offer only a limited array of support services. Some may offer a large menu of support services by subcontracting with external agencies to provide the services. A vendor may house a strong team of software developers and provide sophisticated IT services directly to clients, for example, but subcontract with external consultants when validation research services are required. The in-house expertise of the identified vendor may have important implications for management needs and for ongoing costs associated with use of a particular RNA tool.

### *Costs*

Many costs associated with the use of RNA tools extend beyond the pricing of the instrument itself. Other costs include research, training, software, and other technical assistance services of various forms. For proprietary RNA tools, a batch order for a defined number of assessments may be placed, or a bulk rate may be negotiated per assessment or per case for which assessments and reassessments are to be conducted. Validation research studies and fidelity testing may be included as part of the original service agreement or may be available at an additional cost. Training services are also an additional cost and are typically priced per session. However, most vendors supply train-the-trainer programs to allow local jurisdictions to develop the in-house capacity to conduct future user training sessions. The costs of various software solutions will vary, although ongoing technical assistance support is usually complimentary.

Some RNA instruments are non-proprietary and may be available for use free of charge, but calculations of total cost should consider the availability and pricing of other important support services, such as validation research, fidelity testing, training, and customization of software packages designed for the RNA tool. Some vendors offer support services for the non-proprietary tools reviewed in the Primer's Appendix. If external support is not available or expensive, the jurisdiction should determine whether the costs associated with developing support services or processes of a comparable quality in-house are worth the savings associated with the use of a free RNA tool.

### *Ease of Use*

Finally, the jurisdiction should consider the broader ramifications of adopting a particular RNA tool. This includes considerations related to the user qualifications or requirements to administer the tool. Is the tool complex and difficult to understand? How much staff training is necessary before the tool can be used as compared with other viable options? To administer the LS/CMI, for example, the vendor requires that the staff person: (1) complete a specialized training program administered by an MHS-approved trainer, or (2) document previous completion of graduate-level or professional training on psychometric testing and measurement, or (3) be closely supervised by a test administrator who has completed an approved training program or course.<sup>102</sup>

Another consideration is the amount of staff time involved in proper administration of the tool and use of RNA information. Although the availability of RNA information offers many benefits, the administration of the RNA tool and administrative processes for use of RNA information are often more time consuming than the pre-existing approach. How long will it take to administer



---

and score the assessment? How does the use of RNA information differ from the current approach, and how will the changes in workload affect operations? Will the results be reported in a manner that is easy to incorporate into existing reporting processes, including, for example, to the court? Will the reported information be easy for all users to read, understand, and consistently use in decision-making? These workload efficiency considerations may prompt a need for organizational restructuring. Some jurisdictions, for example, have elected to create a centralized unit in the probation department that is tasked with conducting all initial offender assessments as part of a diagnostic process and producing all presentence investigation reports for the court. In these jurisdictions, supervising probation officers typically conduct subsequent reassessments if the offender is placed on probation.

Finally, the degree of staff support for the use of a RNA tool is also an important consideration. How receptive are judges, staff, and other stakeholders to adoption and implementation of the RNA tool? How committed will they be to using the tool properly and consistently? Greater buy-in from stakeholders may result in more faithful implementation.<sup>103</sup> With guidance from experts, an implementation committee comprised of leadership from local stakeholder groups can be assembled to select an appropriate tool for the jurisdiction.<sup>104</sup> This level of engagement in the initial selection and development process can help to ensure that all stakeholder perspectives are heard at the outset, and can be effective in establishing the necessary foundation of support. In some cases, it may make more sense for a jurisdiction to simply expand the use of an existing RNA tool already employed by the local probation department, if the culture surrounding the use of the RNA tool is a positive one and the tool meets the psychometric standards previously described.

#### *A Note Regarding the Decision to Develop a New Risk and Needs Assessment Tool*

In some cases, jurisdictions may elect to develop, validate, implement, and support the ongoing use of their own RNA tools. Compared with adoption of a RNA tool “off-the-shelf,” this approach requires a larger initial financial investment to support the time-consuming development efforts. The jurisdiction will need to hire professional scientific research personnel with expertise in psychometrics and experience working with criminal justice populations.<sup>105</sup> These researchers should develop a RNA instrument appropriate for use in the jurisdiction, conduct an initial validation study of the new tool, establish a training curriculum for local staff and stakeholders on the proper use of the tool, help establish local capacity to implement the training curriculum in the long term by training local trainers, and provide guidance on the future steps required to maintain the overall effectiveness of the RNA instrument and assessment process over time—including periodic revalidation studies, routine fidelity testing, and other ongoing quality assurance measures. Depending on the research design, the initial validation study of a new RNA instrument alone may take several years to complete.

Because of the time involvement and financial investment associated with developing a new tool, this option may be most advantageous for jurisdictions that already use a RNA tool as part of an established use of evidence-based practices but seek performance improvements such as

---

improved predictive validity or reliability beyond what is perceived possible by using the existing tool and process. In addition, use of a locally developed RNA tool may incur fewer ongoing costs, for example, by eliminating the costs of purchasing a proprietary assessment and by assembling other support services piecemeal, perhaps through a competitive bidding process.<sup>106</sup> Local stakeholders also may feel a greater sense of ownership of the new instrument and process that can, in the long term, stimulate greater support for and more faithful implementation of the tool.

---

## CONCLUSION

---

The proper use of validated, actuarial RNA instruments in assessing the level of risk and criminogenic needs of offenders subject to probation or community supervision is an established evidence-based practice and essential to the success of any serious recidivism reduction enterprise. In this Primer, we have sought to address key questions that judges, probation leaders, and other stakeholders may have about RNA tools in order to assist them in making knowledgeable decisions about the adoption and use of such tools. We have also provided detailed user-friendly information about six commonly used tools in community supervision agencies. Armed with the information provided in this Primer we are confident that criminal justice practitioners will be well-prepared to secure accurate, objective, and reliable risk and needs assessment information on offenders within their jurisdiction.

But even the most accurate, reliable, and fair RNA tool, properly administered by well-trained staff, will not automatically result in changing offender behavior or reducing offender recidivism. A properly validated tool and well-trained officers administering the instrument are certainly two necessary conditions for the effective use of risk and needs assessment information. But much more is also required. Probation officers, judges, and other stakeholders must also be well-trained on other aspects of evidence-based corrections practice: how to use RNA information in tailoring supervision plans and probation orders, how to motivate and effectively supervise offenders to comply with conditions of probation, how to help offenders develop the skills to sustain law-abiding behaviors, and how to most effectively respond to violations of supervision conditions. In addition, sufficient demonstrably effective treatment resources must be available in the community to address offenders' criminogenic needs. Many external providers offer training programs designed to develop and enhance probation skill sets that are critical to effective supervision, and research services to evaluate treatment programs for efficacy.

Accurate assessment is essential but wasted effort unless it leads to effective supervision and treatment. Like assessment and diagnosis in medicine, accurate assessment in corrections is only the first step in the process of developing and then implementing an effective treatment plan. But the fact remains that it is a critical first step: if the initial assessment is inaccurate, the resulting course of supervision and treatment is likely to fail. The authors hope this Primer provides judges and other stakeholders with the information they need to successfully plan and undertake this critical first step in establishing sentencing and community corrections practices that are effective in reducing offender recidivism.

---

## NOTES & REFERENCES

---

<sup>1</sup> Although the last decade has seen an intense focus on identifying programs and practices that work, i.e., evidence-based practices, many researchers and practitioners began exploring the issue much earlier. See Taxman et al. (2013): “Since the early 1990s, the risk-need-responsivity (RNR) model for correctional programming has served as a framework to promote the use of evidence-based correctional strategies” (p. 73). Taxman, F. S., Pattavina, A., Caudy, M. S., Byrne, J., & Durso, J. (2013). The empirical basis for the RNR model with an updated RNR conceptual framework. In F. S. Taxman & A. Pattavina (Eds.), *Simulation strategies to reduce recidivism* (pp. 73-111). New York: Springer. Also see Eisenberg, M., & Markley, G. (1987). Something works in community supervision. *Federal Probation*, 51, 28-32.

<sup>2</sup> See, for example, Andrews, D. A., & Bonta, J. (2006). *The Psychology of Criminal Conduct*, (4th ed.). Cincinnati: Anderson.

<sup>3</sup> See, for example, the National Institute of Justice, Office of Justice Programs, CrimeSolutions.gov website at [www.crimesolutions.gov](http://www.crimesolutions.gov). For information about how programs and practices are rated as effective, see the About CrimeSolutions.gov, Program Review and Rating from Start to Finish website page at [http://www.crimesolutions.gov/about\\_starttofinish.aspx](http://www.crimesolutions.gov/about_starttofinish.aspx). Also see the National Institute of Corrections, Evidence-Based Decision Making website at <http://nicic.gov/ebdm> for information on using evidence-based practices and programs throughout the criminal justice system.

<sup>4</sup> Casey, P. M., Warren, R. K., & Elek, J. K. (2011). *Using offender risk and needs assessment information at sentencing: Guidance for courts from a National Working Group*. Williamsburg, VA: National Center for State Courts. Retrieved from <http://www.ncsc.org/sitecore/content/microsites/csi/home/Topics/~media/Microsites/Files/CSI/RNA%20Guide%20Final.ashx>

<sup>5</sup> The guiding principles expressly state that RNA information should be used for decisions regarding risk reduction and management and not for decisions regarding the appropriate punishment for an offender.

<sup>6</sup> Conference of Chief Justices & Conference of State Court Administrators. (2011). *Resolution 7 in support of the guiding principles on using risk and needs assessment information in the sentencing process*. Williamsburg, VA: National Center for State Courts. Retrieved from <http://ccj.ncsc.org/~media/Microsites/Files/CCJ/Resolutions/o8o32o11-Support-Guiding-Principles-Using-Risk-Needs-Assessment-Information-Sentencing-Process.ashx>

<sup>7</sup> RNA instruments discussed in the Primer are also referred to as “RNA tools” because they include options for generating automated case plans and various management reports.

<sup>8</sup> Some risk assessments are used to classify offenders for various reasons but do not include a needs assessment to identify intervention targets for recidivism reduction. For example, the Virginia Criminal Sentencing Commission (VCSC) developed a risk assessment instrument to identify prison-bound offenders who were low risk to reoffend for purposes of diverting them to a non-prison alternative. See Ostrom, B. J., Kleiman, M., Cheesman, F., Hansen, R. M., & Kauder, N. B. (2002). *Offender risk assessment in Virginia: A three-stage evaluation*. Williamsburg, VA: National Center for State Courts. Retrieved from [http://www.ncsonline.org/WC/Publications/Res\\_Senten\\_RiskAssessPub.pdf](http://www.ncsonline.org/WC/Publications/Res_Senten_RiskAssessPub.pdf). The Primer does not focus on these “risk-only” instruments.

<sup>9</sup> For information on specific offender characteristics, see Andrews et al. (2004): Specific offender characteristics include factors that are not linked to recidivism in the general offender population, “but when they do occur in a given case, they may take a particularly prominent role in the

---

assessment of the offender's risk" (p. 23). Examples of specific offender characteristics are problems with compliance, a diagnosis of psychopathy or other personality disorder, problem-solving and self-management skill deficits, and poor social skills. Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004). *LS/CMI Level of Service/Case Management Inventory: An offender assessment system—user's manual*. North Tonawanda, NY: Multi-Health Systems. Also see Exhibit B "Terms and Definitions from the Risk-Needs-Responsivity Model" in Section 1 "What Are Risk and Needs Assessment Instruments, and Why Use Them?"

<sup>10</sup> See pp. 29-31 in Vincent, G. M., Guy, L. S., Grisso, T. (2012, November). *Risk assessment in juvenile justice: A guidebook for implementation*. Chicago: MacArthur Foundation. Retrieved from <http://www.modelsforchange.net/publications/346>

<sup>11</sup> Casey et al. (2011) at note 4.

<sup>12</sup> Gottfredson, S. D., & Moriarty, L. J. (2006). Clinical versus actuarial judgments in criminal justice decisions: Should one replace the other? *Federal Probation*, 70(2), 15-18. See also, Harris, P. M. (2006). What community supervision officers need to know about actuarial risk assessment and clinical judgment. *Federal Probation*, 70(2), 8-14.

<sup>13</sup> Oleson, J. C., VanBenschoten, S. W., Robinson, C. R., & Lowenkamp, C. T. (2006). Training to see risk: Measuring the accuracy of clinical and actuarial risk assessments among federal probation officers. *Federal Probation*, 75(2), 52-56 at 54.

<sup>14</sup> See Gottfredson & Moriarty (2006) at note 12, p. 15.

<sup>15</sup> This is a generalization of the process. The Appendix describes the specific development process for six commonly used instruments.

<sup>16</sup> See Andrews & Bonta (2006) at note 2, pp. 279-284. Also see, Bonta, J., & Andrews, D. A. (2007). *Risk-need-responsivity model for offender assessment and rehabilitation (PS3-1/2007-6)*. Ottawa: Public Safety Canada. A summary of the RNR model also is included in Casey, et al. (2011) at note 4.

<sup>17</sup> The terms and definitions referenced in Exhibit B are drawn from Andrews & Bonta (2006) at note 2 and Bonta & Andrews (2007) at note 16.

<sup>18</sup> Bonta, J. (2007). Offender Risk Assessment and Sentencing. *Canadian Journal of Criminology and Criminal Justice*, 49, 519-529 at 520. Also see Taxman (2006): "An actuarial-based risk screen is important to determine the degree to which offenders should be given services and resources to ameliorate criminal behavior. The type of services is determined by how the offender "scores" or presents on several criminogenic areas. Those offenders with high criminogenic needs, particularly those that are high or moderate risk, should be given services to ameliorate the criminogenic need, which should reduce the risk for recidivism" (p. 7). Taxman, F. (2006). Assessment with a flair: Offender accountability in supervision plans. *Federal Probation*, 70(2), 2-7.

<sup>19</sup> Andrews & Bonta (2006) at note 2, pp. 73-74. Also see, Andrews, D. A., & Dowden, C. (2007). The risk-need-responsivity model of assessment and human service in prevention and corrections: Crime-Prevention jurisprudence. *Canadian Journal of Criminology and Criminal Justice*, 49, 439-464.

<sup>20</sup> Casey, et al. (2011) at note 4, pp. 7-8.

<sup>21</sup> Vincent et al. (2012, November) at note 10, p. 58.

<sup>22</sup> See Section 1 "What Are Risk and Needs Instruments, and Why Use Them?" for a general description of how a RNA tool is developed; also see the instrument profiles in the Appendix for a description of each tool's specific development process.

<sup>23</sup> Latessa, E., J., Lemke, R., Makarios, M., Smith, P., & Lowenkamp, C. T. (2010). The creation and validation of the Ohio Risk Assessment System (ORAS). *Federal Probation*, 74, 16-22. Retrieved

---

from <http://www.uscourts.gov/viewer.aspx?doc=/uscourts/FederalCourts/PPS/Fedprob/2010-06/index.html>

<sup>24</sup> Barnoski, R., & Drake, E. (2007). Washington's Offender Accountability Act: Department of Corrections' static risk assessment (Doc. No. 07-03-1201). Olympia, WA: Washington State Institute for Public Policy. Retrieved from

[http://www.wsipp.wa.gov/ReportFile/977/Wsipp\\_Washingtons-Offender-Accountability-Act-Department-of-Corrections-Static-Risk-Instrument\\_Full-Report-Updated-October-2008.pdf](http://www.wsipp.wa.gov/ReportFile/977/Wsipp_Washingtons-Offender-Accountability-Act-Department-of-Corrections-Static-Risk-Instrument_Full-Report-Updated-October-2008.pdf)

<sup>25</sup> Baird, C. S., Heinz, R. C., & Bemus, B. J. (1979). *The Wisconsin case classification/staff deployment project: A two-year follow-up report*. Madison, WI: Wisconsin Division of Corrections.

<sup>26</sup> Gottfredson, S., & Moriarty, L. (2006). Statistical risk assessment: old problems and new applications. *Crime & Delinquency*, 52, 178-200. doi: 10.1177/001128705281748. See also Johnson, K. D., & Hardyman, P. L. (2004). How do you know if the risk assessment instrument works? In National Institute of Corrections (Series Ed.), *Topics in Community Corrections: Assessment Issues for Managers* (pp. 20-26). Washington, DC: National Institute of Corrections.

<sup>27</sup> Wright, K. N., Clear, T. R., & Dickson, P. (1984). Universal applicability of probation risk-assessment instruments. *Criminology*, 22, 113-134. doi: 10.1111/j.1745-9125.1984.tb00291.x

<sup>28</sup> See, also, "Is it valid for all offender populations?" in Section 3 "What Are the Qualities of Good Risk and Needs Assessment Instruments?" for a discussion about gender-responsive assessment.

<sup>29</sup> For further discussion of instrument validation, see Section 4 "What Practices Support Sound Implementation of Risk and Needs Assessment Instruments?"

<sup>30</sup> Those instruments that produce a composite risk and needs score (e.g., LSI-R, LS/CMI, ORAS, OST) also provide scores on individual needs domains (e.g., antisocial associates, antisocial attitudes, family and social support), COMPAS provides a needs score for individual domains rather than a total needs score. The CAIS instrument collects information on both risk and needs factors and provides a risk score. It does not provide a total or separate needs score; rather it identifies areas that should be addressed in an offender's case plan. The WRN uses separate instruments to provide a total risk score and a total needs score. STRONG uses separate instruments to provide a total risk score and a needs score for individual domains.

<sup>31</sup> Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior*, 29, 355-379. doi: 10.1177/0093854802029004002. See definitions of Risk-Need-Responsivity terms in Exhibit B in Section 1 "What Are Risk and Needs Assessment Instruments, and Why Use Them?"

<sup>32</sup> B. Lovins, personal communication, February 16, 2012. Latessa et al. (2010) at note 23.

<sup>33</sup> For example, see Vose, B., Smith, P., & Cullen, F. T. (2013). Predictive validity and the impact of change in total LSI-R score on recidivism. *Criminal Justice and Behavior*, 40, 1383-1396. See also Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims & Offenders*, 5, 203-219.

<sup>34</sup> For example, see (1) Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003, January). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, DC: The Institute on Crime, Justice, and Corrections at The George Washington University. (2) Barnoski, R., & Aos, S. (2003). *Washington's Offender Accountability Act: An analysis of the Department of Corrections' risk assessment* (Doc. No. 03-12-1202). Olympia, WA: Washington State Institute for Public Policy. Retrieved from <http://www.wsipp.wa.gov/Reports/03-12-1202>. (3) Caudy, M. S., Durso, J. M., & Taxman, F. S. (2013). How well do dynamic needs predict recidivism? Implications for risk assessment and risk reduction. *Journal of Criminal Justice*, 41, 458-466.

<sup>35</sup> Baird, C. (February, 2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Madison, WI: National Council on Crime and Delinquency. Retrieved from

---

[http://www.nccdglobal.org/sites/default/files/publication\\_pdf/special-report-evidence.pdf](http://www.nccdglobal.org/sites/default/files/publication_pdf/special-report-evidence.pdf). See also Gottfredson & Moriarty (2006) at note 26, p. 192, warning against confusing the purposes of risk and needs assessment:

Although we focus on risk and/or needs assessment as if they can be interchanged, they are very different. We would argue that predicting who will or will not behave criminally is risk assessment, whereas using predictive methods to attempt a reduction in criminality through assignment to differential treatments is needs assessment.

<sup>36</sup> The development of the Wisconsin and STRONG needs assessments, for example, involved input from probation officers regarding factors they considered important to know in preparing case plans. See Baird et al. (1979) at note 25, pp. 12-13. STRONG information is from R. Barnoski, personal communication, April 24, 2012.

<sup>37</sup> For example, the STRONG needs assessment has not been validated. R. Barnoski, personal communication, April 24, 2012. The Wisconsin needs instrument was validated for cases management purposes—not for risk reduction purposes. Researchers and practitioners wanted a tool to better estimate the amount of supervision time a case would require based on the extent of an offender’s problems and deficit areas. The weighted items of the needs assessment provided “a reasonably accurate relationship between the time needed for service delivery and overall need scores.” See Baird et al. (1979) at note 25, p. 14.

<sup>38</sup> Critics also see the risk-only tools as unhelpful for reassessments. Because the risk assessment component of these tools is often comprised of predominantly static risk items (e.g., criminal history), use of the same risk-only assessment after the offender has successfully completed a treatment intervention is unlikely to produce recidivism risk results much different from the offender’s original assessment. The CAIS addresses this issue by using a different risk assessment instrument for reassessment. C. Baird, personal communication, July 24, 2012. For the two different versions of the risk assessment instruments, see National Council on Crime and Delinquency. (2010). *CAIS Correctional Assessment and Intervention System: System manual*. Madison, WI: Author.

<sup>39</sup> See Skeem & Monahan commenting on the value of composite versus separate risk and needs instruments, assuming all have been validated:

If the ultimate purpose is to manage or reduce an individual’s risk, then value may be added by choosing an instrument that includes treatment-relevant risk factors. (Although an integrated instrument would be most parsimonious, we can easily envision a two-stage process in which a risk assessment step was followed by an independent risk management step.) This choice is appropriate for ongoing decisions in which the risk estimate can be modified to reflect ebbs and flows in an individual’s risk over time. Beyond focusing risk reduction efforts, these instruments could provide incentive for changing behavior (a parole board cannot advise an inmate to undo his past commission of an assault but can advise him to develop employment skills).” (p. 41)

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, 20(1), 38-42. doi: 10.1177/0963721410397271

<sup>40</sup> Information on each item’s relationship to recidivism was not found in the studies reviewed for the COMPAS. The COMPAS *Practitioner’s Guide* provides information on the relationship between its risk scales and recidivism. See pp. 17-19 in Northpointe Institute for Public Management. (2013, January). *Practitioner’s guide to COMPAS*. Traverse City, MI: Author. Retrieved from

[http://www.northpointeinc.com/files/technical\\_documents/FieldGuide2\\_012813.pdf](http://www.northpointeinc.com/files/technical_documents/FieldGuide2_012813.pdf)

---

<sup>41</sup> Caudy et al. (2013) at note 34, p. 465: “In order for risk reduction strategies to be effective, needs assessments must be validated and linked to specific evidence-based interventions.”

<sup>42</sup> Vincent et al. (2012, November) at note 10, p. 34. See, also, Andrews & Bonta (2006) at note 2, p. 48.

<sup>43</sup> See, for example, Andrews, D.A., Bonta, J. L., & Wormith, J. S. (2009). *Level of Service/Case Management Inventory (LS/CMI) supplement: A gender-informed risk/need/responsivity assessment*. Toronto, ON: Multi-Health Systems, Inc.

<sup>44</sup> Bonta & Andrews (2007) at note 16. See, also, Crime and Justice Institute. (2004). *Implementing evidence-based principles in community corrections: The principles of effective intervention*. Boston: Author. Retrieved from <http://static.nicic.gov/Library/019342.pdf>

<sup>45</sup> For the LS/CMI, strengths are not included in the quantitative risk and need score. See Andrews et al. (2004) at note 9, p. 4. For information on the OST, see Arizona Adult Probation Services Division. (2009, July update). *OST Scoring Guide*. AZ: Authors.

<sup>46</sup> Because a risk-only instrument typically includes many static risk items, it is more quickly and easily scored. A probation department can screen an offender for risk right away and then conduct a needs assessment at a later date to inform case planning decisions. This approach also allows a jurisdiction to use the risk-only assessment to triage assessment administration resources. For example, if an offender is determined to be low-risk and therefore not an appropriate target for intensive risk-reduction treatment services, probation may determine that a full needs assessment is unnecessary whereas offenders determined to be moderate- or high-risk would be given a needs assessment. Composite risk and needs assessment instruments address this issue by providing or recommending a separate “quick screen” tool. See, for example, the ORAS Community Supervision Screening Tool, pp. 29-31 in Latessa, E. J., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. T. (2009, July). *The creation and validation of the Ohio Risk Assessment System: Final report*. Cincinnati: Center for Criminal Justice Research, University of Cincinnati School of Criminal Justice. Retrieved from [http://www.ocjs.ohio.gov/ORAS\\_FinalReport.pdf](http://www.ocjs.ohio.gov/ORAS_FinalReport.pdf). A brief risk-only assessment may also be used “for expedited or early disposition cases to provide additional information to the court that otherwise would not be available because the person did not go through the presentence investigation process.” See p. 1 in Arizona Adult Probation Services Division (2009, July update). *MOST Scoring Guide*. AZ: Authors. Vincent et al. (2012, November) at note 10, pp. 58-59 caution that such screening tools should be used when risk is the only question; they should not be used to guide treatment planning.

<sup>47</sup> Some RNA instruments (e.g., LSI-R, LS/CMI, COMPAS) also provide information on another form of reliability referred to as *internal consistency*. Internal consistency reliability provides an indication of the extent to which all the items in a scale measure the same single underlying concept or dimension. The test commonly used to measure internal consistency is called Cronbach’s alpha. Because RNA tools are deliberately designed to measure multiple multifaceted factors related to recidivism rather than a single construct, test developers generally focus more on the tool’s predictive accuracy than on its internal consistency.

<sup>48</sup> See p. 286 in Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284-290.

<sup>49</sup> Other statistical techniques also capture inter-rater reliability, but are less commonly used in existing RNA research. The OST and WRN profiles report studies using percent agreement between raters.



---

<sup>50</sup> For more information about training and quality assurance, see Section 4 “What Practices Support Sound Implementation of Risk and Needs Assessment Instruments?”

<sup>51</sup> Examples of other forms of validity not covered in this Primer are **content validity**, or the degree to which the RNA tool measures all of the information that is conceptually relevant to a complete understanding of recidivism risk; **face validity**, or the degree to which the instrument makes intuitive sense to probation officers and other stakeholders (which can be important in motivating staff to actually use the tool); and **concurrent validity**, or the degree to which a new RNA tool reflects the same constructs measured by an existing or “gold standard” RNA tool.

<sup>52</sup> Grove, W., & Meehl, P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2, 293-323 at 306.

<sup>53</sup> For more on this issue, see “*Use of administrative overrides*” in Section 4 “What Practices Support Sound Implementation of Risk and Needs Assessment Instruments?”

<sup>54</sup> See Gottfredson & Moriarty (2006) at note 12; Grove & Meehl (1996) at note 52; Latessa & Lovins (2010) at note 33; and Skeem & Monahan (2011) at note 39.

<sup>55</sup> Vincent (November 6, 2012) at note 10, pp. 81-82.

<sup>56</sup> See note 55.

<sup>57</sup> Researchers may report other statistics, such as Pearson’s chi-square ( $\chi^2$ ) test to determine how much separation an assessment tool achieves between risk level classifications or contingency tables with a Relative Improvement Over Chance (RIOC) value to quantify how much improvement the tool introduces over chance. However, these statistical techniques are less commonly reported.

<sup>58</sup> When creating or revalidating an RNA tool, researchers will often examine whether each item is related to recidivism, how each item in the assessment is weighted before the item scores are summed to create a raw recidivism risk score, and at which points in the continuum of raw risk scores could cutoffs be introduced to define the low, moderate, and high risk level classifications. For example, responses on each item in the original Wisconsin risk assessment determined the item’s score. After the item scores are summed to create a raw risk score, cutoff values of 8 and 15 were used to create the low, moderate, and high risk classification groups. See Baird et al. (1979) at note 25, p. 11. See, also, Baird (2009) at note 35, pp. 6-7, discussing the importance of examining recidivism rates by risk level in evaluating a risk assessment system.

<sup>59</sup> Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, 63, 737-748.

<sup>60</sup> See, for example, (1) Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, 62, 783-792. (2) Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1-26. (3) Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen’s d, and r. *Law and Human Behavior*, 29, 615-620.

<sup>61</sup> The guidelines are based on the following works of Rice and Harris: Rice & Harris (1995) at note 59 and Rice & Harris (2005) at note 60. See, also, Hanson, R. K. (January, 2000). Risk assessment. Beaverton, OR: Association for the Treatment of Sexual Abusers. Retrieved from [http://www.cj-resources.com/CJ\\_Corrections\\_pdfs/InfoPac%20Risk%20assessment%20booklet%20-%20Hanson%202000.pdf](http://www.cj-resources.com/CJ_Corrections_pdfs/InfoPac%20Risk%20assessment%20booklet%20-%20Hanson%202000.pdf)

<sup>62</sup> Gendreau, P., Little, T., & Goggin, C. E. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34, 575-607. doi: 10.1111/j.1745-9125.1996.tb01220.x

---

<sup>63</sup> For example, as suggested by Hanson, even small effect sizes “may have considerable consequences in some contexts” (p. 176). Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology*, 50, 172-182.

<sup>64</sup> Hanson (2009) at note 63.

<sup>65</sup> See Rice & Harris (2005), at note 60, p. 619 and Hanson (2009) at note 63, p. 176.

<sup>66</sup> Researchers often examine subgroupings by particular offender demographic or descriptive characteristics like gender or race but may also examine differential validity by type of offense committed (e.g., among felony property offenders, felony drug offenders). Risk of violent crime reoffending and sex crime reoffending are often of particular interest to leaders and policymakers in the criminal justice system, but general risk assessment instruments typically are not developed and validated to address these specific forms of recidivism. Instead, specialized assessment tools have been developed specifically for estimating the likelihood that an offender will commit another violent crime or sex crime.

<sup>67</sup> Van Voorhis, P., Salisbury, E. J., Wright, E. M., & Bauman, A. (2008). *Achieving accurate pictures of risk and identifying gender responsive needs: Two new assessments for women offenders*. Washington, DC: National Institute of Corrections. Retrieved from

<http://www.uc.edu/content/dam/uc/womenoffenders/docs/NIC%20Summary%20Report.pdf>

See also Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women’s risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior*, 37, 261-288. doi: 10.1177/0093854809357442

<sup>68</sup> Examples of criminogenic needs for women are parental stress, family support, anger, depression and other symptoms of mental illness, unsafe housing, educational assets, self-esteem, and self-efficacy. See Van Voorhis et al. (2008) at note 67, p. 14.

<sup>69</sup> For the LS/CMI, see Andrews et al. (2009) at note 43. For the COMPAS, see Brennan, T., Breitenbach, M., & Dieterich, W. (2008). *A need/risk explanatory classification of female prisoners incorporating gender-neutral and gender-responsive factors*. Traverse City, MI: Northpointe Institute for Public Management. Retrieved from

[http://www.northpointeinc.com/files/research\\_documents/A\\_Need-Risk\\_Explanatory\\_Classification\\_of\\_Females.pdf](http://www.northpointeinc.com/files/research_documents/A_Need-Risk_Explanatory_Classification_of_Females.pdf)

<sup>70</sup> A good tool often incorporates information from multiple methods of data collection, as this often results in gains in predictive validity. See Bonta (2002) at note 31.

<sup>71</sup> Northpointe Institute for Public Management (2013, January) at note 40, pp. 44-45.

<sup>72</sup> COMPAS also includes a Random Responding scale to identify offenders who may be randomly answering the questionnaire. See note 71.

<sup>73</sup> When choosing among existing RNA tools, some researchers recommend selecting a tool that has been evaluated by independent researchers in at least two separate studies. See Vincent et al. (November 6, 2012) at note 10.

<sup>74</sup> See p. 502 in Lowenkamp, C. T., & Latessa, E. J. (2005). Increasing the effectiveness of correctional programming through the risk principle: Identifying offenders for residential placement. *Criminology & Public Policy*, 4, 263-290. doi: 10.1111/j.1745-9133.2005.00021.x

<sup>75</sup> Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. In National Institute of Corrections (Series Ed.), *Topics in Community Corrections: Assessment Issues for Managers* (pp. 3-8). Washington, DC: National Institute of Corrections.

<sup>76</sup> Johnson & Hardyman (2004) at note 26.

- 
- <sup>77</sup> See p. 479 in Ferguson, J. L. (2002). Putting the "What Works" research into practice: An organizational perspective. *Criminal Justice and Behavior*, 29, 472-492. Retrieved from <http://cjb.sagepub.com/content/29/4/472>.
- <sup>78</sup> White, T. F. (2004). Implementing an offender risk and needs assessment: An organizational change process. In National Institute of Corrections (Series Ed.), *Topics in Community Corrections: Assessment Issues for Managers* (pp. 42-48). Washington, DC: National Institute of Corrections.
- <sup>79</sup> Vincent et al. (November 6, 2012) at note 10, pp. 81-82.
- <sup>80</sup> That is, any RNA tool that generates a score for categorization purposes or which reports the probability of recidivism as a ratio or percentage likelihood. See note 79.
- <sup>81</sup> See p. 3 in Assessments.com. (2009). *The STRONG: Static Risk and Offender Needs Guide*. Bountiful, UT: Authors. Retrieved from [http://www.assessments.com/assessments\\_documentation/STRONG%20Fact%20Sheet.pdf](http://www.assessments.com/assessments_documentation/STRONG%20Fact%20Sheet.pdf)
- <sup>82</sup> B. Lovins, personal communication, February 16, 2012. R. Barnoski, personal communication, April 24, 2012.
- <sup>83</sup> Note that the independent researchers conducting this revalidation study also strongly recommended removal of an item that the original instrument developers acknowledged was not associated with recidivism but included in the original risk assessment instrument solely for policy reasons. Eisenberg, M., Bryl, J., and Fabelo, T. (July, 2009). *Validation of the Wisconsin Department of Corrections Risk Assessment Instrument*. New York: Council of State Governments Justice Center. Retrieved from <http://csgjusticecenter.org/wp-content/uploads/2012/12/WIRiskValidationFinalJuly2009.pdf>
- Baird et al. (1979) at note 25 and C. Baird, personal communication, July 24, 2012.
- <sup>84</sup> Clear, T. R., & Gallagher, K. W. (1985). Probation and parole supervision: A review of current classification practices. *Crime & Delinquency*, 31, 423-443. doi: 10.1177/001128785031003007
- <sup>85</sup> For example, one study confirmed the predictive validity of the LSI-R, but only when the assessment was scored by staff formally trained on how to properly administer the assessment. The relationship between LSI-R results and recidivism disappeared when untrained staff administered the tool. See Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity. *Journal of Criminal Justice*, 34, 523-529. doi: 10.1016/j.jcrimjus.2006.09.007
- <sup>86</sup> See *Guiding Principle 4: Stakeholder Training* (pp. 21-22) in Casey et al. (2011) at note 4.
- <sup>87</sup> Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida. Retrieved from <http://ctndisseminationalibrary.org/PDF/nirnmonograph.pdf>
- <sup>88</sup> Joyce, B., & Showers, B. (2002). *Student achievement through staff development* (3<sup>rd</sup> ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
- <sup>89</sup> Vincent et al. (2012, November) at note 10, p. 86.
- <sup>90</sup> Fixsen et al. (2005) at note 87.
- <sup>91</sup> See the STRONG instrument profile in the Appendix.
- <sup>92</sup> Refer to the *Override Policy* sections for each RNA tool featured in the Appendix.
- <sup>93</sup> B. Lovins, personal communication, December 7, 2012.
- <sup>94</sup> Grove & Meehl (1996) at note 52. See, also, Gottfredson & Moriarty (2006) at note 26.
- <sup>95</sup> See, for example, Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision*
-

---

*Processes*, 37, 93-110. Also, Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgment. *American Psychologist*, 23, 483-496.

<sup>96</sup> Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, 39, 1511-1538. doi: 10.1177/0093854812455741

<sup>97</sup> See, for example, Center for Sentencing Initiatives (December, 2013). *Use of risk and needs assessment information at sentencing: Napa County, California*. Williamsburg, VA: National Center for State Courts. Retrieved from

<http://www.ncsc.org/sitecore/content/microsites/csi/home/Topics/~media/Microsites/Files/CSI/RNA%20Brief%20-%20Napa%20County%20CA%20ocsi.ashx>

Center for Sentencing Initiatives (December, 2013). *Use of risk and needs assessment information at sentencing: Grant County, Indiana*. Williamsburg, VA: National Center for State Courts. Retrieved from

<http://www.ncsc.org/sitecore/content/microsites/csi/home/Topics/~media/Microsites/Files/CSI/RNA%20Brief%20-%20Grant%20County%20IN%20ocsi.ashx>

<sup>98</sup> See, for example, Center for Sentencing Initiatives (December, 2013). *Use of risk and needs assessment information at sentencing: Mesa County, Colorado*. Williamsburg, VA: National Center for State Courts. Retrieved from

<http://www.ncsc.org/sitecore/content/microsites/csi/home/Topics/~media/Microsites/Files/CSI/RNA%20Brief%20-%20Mesa%20County%20CO%20ocsi.ashx>

Center for Sentencing Initiatives (December, 2013). *Use of risk and needs assessment information at sentencing: 7<sup>th</sup> Judicial District, Idaho*. Williamsburg, VA: National Center for State Courts. Retrieved from

<http://www.ncsc.org/sitecore/content/microsites/csi/home/Topics/~media/Microsites/Files/CSI/RNA%20Brief%20-%207th%20Judicial%20District%20ID%20ocsi.ashx>

<sup>99</sup> Kreamer, S. (2004). Quality assurance and training in offender assessment. In National Institute of Corrections (Series Ed.), *Topics in Community Corrections: Assessment Issues for Managers* (pp. 13-19). Washington, DC: National Institute of Corrections.

<sup>100</sup> Vincent et al. (2012, November) at note 10, p. 84.

<sup>101</sup> See the ORAS instrument profile in the Appendix.

<sup>102</sup> Andrews et al. (2004) at note 9, p. 6.

<sup>103</sup> Ferguson (2002) at note 77.

<sup>104</sup> Vincent et al. (2012, November) at note 10, p. 57.

<sup>105</sup> Vincent et al. (2012, November) at note 10. See, also, Del Pra, Z. (2004). In search of a risk instrument. In National Institute of Corrections (Series Ed.), *Topics in Community Corrections: Assessment Issues for Managers* (pp. 9-12). Washington, DC: National Institute of Corrections.

<sup>106</sup> Ferguson (2002) at note 77.

---

## APPENDIX

---

### *Risk and Needs Assessment Instrument Profiles*

This Appendix reviews six risk and needs assessment (RNA) tools. As explained in the Primer, each profile begins with a glossary of definitions for common terms used in the creation of the RNA tools. The terms and their definitions vary somewhat across the tools. The profiles also present information on the following general categories: (a) history and current use, (b) development, (c) content, (d) instrument reliability and validity, and (e) practical considerations. The profiles are based on a review of the literature and interviews with at least one individual involved in the development of each instrument. The instrument developers also had an opportunity to respond to a discussion guide prepared for each instrument that was revised following each interview as well as the final draft versions of the profiles.

Readers are encouraged to read the Primer to gain a broader context regarding the purpose and appropriate use of RNA tools and a better understanding of some of the terms (e.g., reliability and validity) used in the profiles.

<i>RNA Instrument Profile</i>	<i>Page</i>
Correctional Assessment and Intervention System (CAIS)	A-2
Correctional Offender Management Profile for Alternative Sanctions (COMPAS)	A-19
Level of Service Assessments (LSI-R and LS/CMI)	A-30
Offender Screening Tool (OST)	A-44
Ohio Risk Assessment System (ORAS)	A-52
Static Risk and Offender Needs Guide (STRONG)	A-59

# Correctional Assessment And Intervention System (CAIS)\*

---

### CAIS GLOSSARY OF TERMS

---

<b><i>Risk</i></b>	Risk refers to the aggregate likelihood that an offender classified into a particular risk group will commit subsequent criminal behavior. <sup>1</sup>
<b><i>Static risk</i></b>	Christopher Baird, a National Council on Crime and Delinquency (NCCD) CAIS author, recognizes the use of static (i.e., not changeable) and dynamic (i.e., changeable) factors by the field but did not use those distinctions in developing the instrument: “Any factor (other than those that should not be included for ethical reasons) that adds to the instrument’s ability to optimally separate risk groups should be included in a risk tool. It does not matter if a factor is static or dynamic.” <sup>2</sup>
<b><i>Dynamic risk</i></b>	See above. Baird describes the CAIS system in its entirety as dynamic. “At reclassification, the emphasis shifts from prior criminal history items to measures that reflect adjustment during supervision,” allowing “clients to move between supervision levels based on their performance.” <sup>3</sup>
<b><i>Needs</i></b>	Needs refer to “problems and deficit areas” most commonly evidenced in probationers and parolees. <sup>4</sup> According to Baird, a particular need is not criminogenic (i.e., causing criminal behavior) in and of itself; rather a need can only be deemed criminogenic for an individual offender. <sup>5</sup>
<b><i>Responsivity</i></b>	Term not used explicitly in reports on the creation of CAIS.
<b><i>Protective factors</i></b>	Term not used in reports on the creation of CAIS. Baird contends that protective factors can be important to case planning and management but is critical of the manner in which these factors have been assessed and used by the field. <sup>6</sup>
<b><i>Strengths</i></b>	CAIS considers strengths and needs in developing supervision strategies. Potential strengths are areas rated by the interviewer as having no or only minor significance in generating criminal behavior. <sup>7</sup>
<b><i>Recidivism</i></b>	The CAIS manual defines recidivism as “the likelihood that an offender will experience a subsequent felony conviction or be revoked into an institutional setting in the next 24 months.” <sup>8</sup>

---

\*The CAIS is based on components of the National Institute of Corrections’ Model Probation and Parole Management Program, including the Wisconsin Risk and Needs (WRN) assessment instruments and the Client Management Classification (CMC) planning guide. Accordingly, these also are discussed in the profile.

## Appendix: RNA Instrument Profile for the CAIS

---

### HISTORY & CURRENT USE.

---

**Creation.** The Correctional Assessment and Intervention System (CAIS) evolved from efforts in Wisconsin, beginning in 1975 at the direction of the state legislature, to develop a case classification system for probationers and parolees that would improve the effectiveness of service delivery.<sup>9</sup> Though the Wisconsin effort began with funding from the Law Enforcement Assistance Administration (LEAA), it required four years and substantial additional resources from the Wisconsin Division of Corrections, Bureau of Community Corrections to design, implement and evaluate.<sup>10</sup> The classification system that emerged from this effort, commonly referred to as the Wisconsin Risk and Needs (WRN) assessment, has separate risk assessment and needs assessment components, each developed independent of the other using different methodologies.<sup>11</sup> The risk and needs scores were used principally to determine an appropriate level of supervision for an offender but did not address case planning and supervision. To address this gap, the Client Management Classification (CMC) system was developed.<sup>12</sup> The CMC uses information about offender needs, as well as other factors thought to distinguish different types of offenders, to classify an offender into one of four supervision categories.

The CAIS combines updated versions of the Wisconsin risk, needs, and supervision strategy assessments into a single, automated system to assist case managers with the effective and efficient supervision of offenders.<sup>13</sup> CAIS provides this information through a web-based data system accessible

via internet browser. In addition to providing individual offender assessment reports, CAIS also has the capability to produce aggregate, managerial reports to help identify service gaps and target resources.<sup>14</sup> Much of the information and research available is on earlier versions of the various CAIS components. Thus this profile reviews the development and application of the WRN and CMC as the precursors to the CAIS.

**Current use.** Numerous correctional agencies outside of Wisconsin adopted the WRN (or a slight variation of the instrument) and the CMC after the instruments became part of the National Institute of Corrections (NIC) Model Probation and Parole Project in the 1980s.<sup>15</sup> A survey of 288 state and local probation and parole agencies by the University of Cincinnati in 1998-1999 reported that the most widely used instrument is the CMC system (36%), including both the WRN instruments; another 26.3% reported using the WRN assessment but not the CMC; 2% reported using only the Wisconsin risk assessment, and less than 1% reported using the Wisconsin needs assessment alone.<sup>16</sup> In addition, inspection of the instruments falling in the “other” category also revealed that some of these instruments were versions of the WRN assessments.

The National Survey of Criminal Justice Treatment Practices, a survey of prisons, jails, and community correctional agencies begun in 2002, identified the WRN as the second most frequently used assessment instrument by these agencies, though the percentage was only 12.7% because nearly two-thirds of the facilities reported not using

## Appendix: RNA Instrument Profile for the CAIS

---

any instrument.<sup>17</sup> It is not known how well these figures reflect current use of the instruments.

CAIS, more recent in its development, is currently used by ten agencies, including county probation departments, a county jail with an associated reentry program, a county reentry program and a non-governmental community-based reentry program.<sup>18</sup>

### DEVELOPMENT.

---

***Instrument purpose.*** CAIS “is a supervision strategy model that weaves together a risk assessment and a needs assessment.”<sup>19</sup> CAIS identifies the underlying motivation for an offender’s criminal behavior to assist in developing the offender’s case plan.

According to its developers, its purpose is to assist case managers with supervising offenders effectively and efficiently with the goals of aiding institutional adjustment, reducing recidivism, and helping offenders live productively in the community.<sup>20</sup>

***Approach to instrument development.***

CAIS is designed to accommodate a variety of risk assessment instruments, but the default instrument is a modified version of the Wisconsin Department of Corrections risk assessment instrument (sometimes referred to as the DOC-502 risk scale).<sup>21</sup> The Wisconsin risk assessment was developed using a criterion variable that combined the number of occurrences of absconsions, rules violations, arrests, misdemeanor convictions, felony convictions, and convictions for assaultive offenses.<sup>22</sup> Utilizing a retrospective design, information was collected on approximately 250 randomly selected closed or revoked cases. A working committee of

probation officers, supervisors and research staff identified 22 items they associated with offender recidivism based on professional judgment and consensus opinion.

Researchers then applied linear regression techniques to refine this pool of items and eliminate items that failed to demonstrate a statistically significant relationship with recidivism. Seven items were retained as a result of this process. To enhance predictive validity, researchers added three items that were not identified by the regression analysis but nonetheless had a strong relationship with the outcome measure (examining item significant differences and simple correlation coefficients) and discriminated among high, moderate, and low risk offenders.<sup>23</sup> The final scale consisted of these ten items, each weighted based on its correlation with criminal behavior.

At the explicit request of the Wisconsin Department of Corrections, the test developers added an eleventh item, history of assaultive offense, to the instrument. The purpose of this item was to ensure that offenders “who had committed an assaultive offense within the last five years are placed under maximum supervision for (at least) the first six months of probation or parole.”<sup>24</sup> The item added 15 points to an offender’s risk assessment score, the minimum score needed to be placed under maximum supervision. At reevaluation, supervision levels were based solely on risk and needs scores; the additional points were not added to the offender’s reevaluation score.<sup>25</sup> The additional assaultive item was never considered to be part of the ten-item actuarial risk scale because it was never shown to be related to the risk of recidivism.<sup>26</sup> However, the item was



## Appendix: RNA Instrument Profile for the CAIS

---

included in some subsequent use of the risk scale by others despite its lack of predictability in the development of the instrument.

After construction, the risk scale was initially tested on a sample of 4,231 Wisconsin offenders. The results indicated that initial risk scores were related to subsequent revocations: Approximately 2% of low risk offenders, 9% of moderate risk offenders, and 26% of high risk offenders were revoked.<sup>27</sup>

The needs component of the WRN assessment, and subsequently the CAIS, was designed to assess the extent of an offender's problems and deficit areas to better estimate the amount of supervision time the case would require.<sup>28</sup> The Wisconsin project also sought to standardize the needs information collected across probation officers.<sup>29</sup>

To develop the needs tool, probation and parole officers and researchers identified an extensive list of possible client needs and, using the list, surveyed incoming clients over an eight-month period in Madison. A set of eleven areas of needs emerged from this process: 1) academic/vocational skills, 2) employment problems, 3) financial management, 4) marital/family relationships, 5) companions, 6) emotional stability, 7) alcohol use, 8) other drug use, 9) mental ability, 10) health, and 11) sexual behavior.<sup>30</sup> Together, these areas were "thought to encompass the wide range of problems that are most commonly evidenced in probationers and parolees."<sup>31</sup>

Each of the eleven items and a twelfth item assessing the probation officer's impression of the offender's needs is weighted based on

supervision time to address the need. Initially based on the professional judgment of the probation and parole agents, the weights were subsequently empirically verified on a sample of 482 offenders as presenting "a reasonably accurate relationship between the time needed for service delivery and overall need scores."<sup>32</sup>

In Wisconsin, agencies used the highest score of either the risk or needs scale to determine the level of supervision.<sup>33</sup> When other states began using the instrument, this practice varied with some states relying more on one or the other instrument—usually the risk assessment.<sup>34</sup> Eventually, most users settled on the risk assessment for determining level of supervision, as is the approach taken with the CAIS.<sup>35</sup>

Once the level of supervision is known, probation officers turned to CMC to develop a case plan and supervision strategy for an offender.<sup>36</sup> The CMC was developed by two clinical psychologists, a line officer, and research staff.<sup>37</sup> The development team began by identifying items with a potential for differentiating among basic offender types. They used the items to create an instrument based on forced-choice ratings, i.e., each item has several possible choices, and the interviewer selects the choice that best describes the offender. To increase the reliability of ratings, the team developed a 45-minute semi-structured interview with scripted questions and a companion scoring guide. The development process eventually yielded 45 offender attitude questions, 11 objective background and offense history items, 8 interview behavior items, and 7 interviewer impression items.

## Appendix: RNA Instrument Profile for the CAIS

---

The development team identified four supervision strategies based on their extensive experience working with offenders.<sup>38</sup> The team assessed a sample of offenders and, based on the assessment, subjectively placed each offender into one of the supervision strategies. The CMC items were then tested to see how strongly each influenced the professionals' decisions.<sup>39</sup> Weights were assigned to each item based on its ability to discriminate among the supervision strategy groups and its interrater reliability score.<sup>40</sup>

The test developers tracked 250 offenders in both the construction and cross-validation samples for 12 months to determine if offender behaviors were consistent with the expected problems and needs associated with the supervision strategy to which they were assigned.<sup>41</sup> The CMC system was modified to improve its reliability and validity based on the resulting data.<sup>42</sup> Using the data and their knowledge of supervision strategies, the test developers created supervision guidelines for the offenders in each strategy group. The guidelines provided information on "offender goals, officer/offender relationships, appropriate auxiliary services and programs, and supervision techniques."<sup>43</sup>

In explaining the development of the CAIS, the National Council on Crime and Delinquency (NCCD) noted that probation and parole agencies had become discontented with the CMC because it was not automated.<sup>44</sup> As a result, NCCD embarked on a two-year process to update and automate the CMC, resulting in the CAIS. CAIS incorporates the Wisconsin risk scale or other validated actuarial risk

assessment, needs items and CMC items.<sup>45</sup> Thus information collected in one interview provided probation officers with an offender's risk level, suggested supervision strategy, and principal service needs. In creating the CAIS, a few items were added (e.g., "What was your behavior like?" was added as a follow-up to "How would you describe yourself as a child?") or revised (e.g., "How much socializing do you do with women (men)?" revised to "Can you tell me about your relationships with women/men?"). In addition, several items (e.g., "Do you have any children?" and "How do you feel about being a mom?") were added for assessment of female offenders. NCCD reports that CAIS developers relied on an expert in gender issues to help develop gender-specific supervision strategies that focus on programs shown to be effective with female offenders.<sup>46</sup> As a result, the supervision and case planning recommendations may be somewhat different than what came out of the original system.<sup>47</sup>

### CONTENT.

---

**Structure.** CAIS generates a report that consists of two sections: Primary Case Planning Approach and the Specific Client Profile.<sup>48</sup> The Primary Case Planning Approach section has five sub-sections: (1) classification (providing scores for each supervision strategy and identifying the primary strategy to follow); (2) general issues facing offenders in the selected strategy; (3) goals of supervision; (4) common needs/referrals for offenders in the supervision strategy; (5) caseworker/offender relationship (providing guidance for working

## Appendix: RNA Instrument Profile for the CAIS

with offenders in the particular supervision strategy); and (6) techniques of supervision (i.e., those that are particularly applicable for the specific supervision strategy).<sup>49</sup>

The Specific Client Profile consists of three sections: (1) risk level; (2) principal service needs; and (3) special concerns.

**Items and domains.** Table 1 summarizes the number of items for each of the four major sections of the CAIS instrument as administered to female and male offenders.

**Table 1. CAIS Sections**

CAIS Sections	# of Items	
	Women	Men
1. General Information		
• Offense Patterns	8	8
• School Adjustment	5	5
• Vocational and Residential Adjustment	7	7
• Family Information	19	17
• Interpersonal Relations	7	7
• Feelings	6	6
• Plans and Problems	5	5
2. Objective History	11	11
3. Behavioral Observations	8	8
4. Interviewer Impressions	12	8
Total Number of Items	88	82

The 11 risk items are embedded within the “General Information” and “Objective History” sections. They are the same for female and male offenders. As noted earlier, however, jurisdictions can opt to replace the default CAIS risk assessment with their own

validated risk instrument if they prefer. The risk items for the CAIS and the original Wisconsin risk instrument are compared in Table 2.<sup>50</sup>

**Table 2. CAIS and Wisconsin Risk Items<sup>51</sup>**

CAIS Risk Items	WRN Risk Items
1. Employment	1. % of time employed in last 12 months
2. Address changes in the last year	2. Address changes in last 12 months
3. Offender’s pattern of associates	
4. Age at first arrest	3. Age at first conviction
5. # of prior offenses	4. # of prior felony convictions
6. Ever convicted for theft, burglary, auto theft, robbery	5. Convictions for burglary, theft, auto theft, robbery, worthless checks or forgery
7. # of prior jail sentences	
8. # of prior periods of probation or parole supervision	6. # of prior periods of probation/parole supervision
9. Ever had probation or parole revoked	7. # of prior probation/ parole revocations
10. % of criminal behavior related to alcohol abuse	8. Alcohol usage problems
11. % of criminal behavior related to other drug use	9. Other drug usage problems
	10. Attitude

Unlike the original Wisconsin needs assessment instrument, the CAIS does not provide an overall need score; rather it identifies areas that should be addressed in the offender’s case plan.

---

## Appendix: RNA Instrument Profile for the CAIS

---

Following the CMC approach, CAIS classifies offenders into one of four supervision strategies. According to the CAIS manual, the classification is based on items from all sections of the CAIS.<sup>52</sup> The manual also explains that “scores for the strategy groups are the result of a complex set of research-based scoring rules.”<sup>53</sup> The four supervision groups are:<sup>54</sup>

- The Selective Intervention (SI) strategy, which includes different strategies for situational (SI-S) and treatment (SI-T) groups, is for offenders who generally have pro-social values, positive adjustment, positive achievements, and good social skills.
- The Casework/Control (CC) strategy is for offenders with a broad range of instability, a chaotic lifestyle, emotional instability, multi-drug abuse/addiction, and negative attitudes towards authority.
- The Environmental Structure (ES) strategy is for offenders who lack social and survival skills, have poor impulse control, are gullible and naïve, and show poor judgment.
- The Limit Setting (LS) strategy is for offenders with antisocial values, who prefer to succeed outside the rules/law, whose role models operate outside the rules/law, and are manipulative and exploitive.

**Reporting risk levels.** The CAIS groups offenders into three levels of risk: low, moderate and high. The CAIS provides initial ranges of scores for each risk level; however, the NCCD, which holds the copyright to the CAIS, reports that “as part of each CAIS implementation project, NCCD validates the risk instrument periodically and customizes

the instrument for each agency to ensure it optimally classifies cases.”<sup>55</sup> NCCD encourages agencies to collect reassessment data which provides information on the current status of a case to assist with validation.<sup>56</sup>

---

### INSTRUMENT RELIABILITY AND VALIDITY.

---

NCCD does not indicate whether the updated and gender-specific versions of the CAIS were evaluated independently. The studies cited in support of the CAIS, and presented in the following sections, are those based on the original CMC.

In addition, Baird describes the Wisconsin risk and needs scales as providing an approach to assessing offender risk and needs.<sup>57</sup> The intent was to provide templates that jurisdictions could customize for their particular populations based on their own validation studies. As a result, there are many versions of the risk and needs scales with minor variations, which should be taken into consideration when comparing the results of validation studies across jurisdictions.

**Populations studied.** In addition to the statewide Wisconsin construction and validation samples of probation-eligible male and female adult offenders, the Wisconsin risk and needs assessment instruments and the CMC have been implemented and studied in a variety of states and Canada.

**Predictive validity.** Gendreau and his colleagues reported a mean effect size of  $r=.27$  between the Wisconsin risk scale and measures of recidivism.<sup>58</sup> The meta-analysis

## Appendix: RNA Instrument Profile for the CAIS

---

was based on 14 effect sizes calculated from various studies. It is not known how many of the effect sizes were based on the instrument with the assaultive factor included versus excluded.<sup>59</sup> Bonta reported correlations of  $r=.22$  to  $r=.33$  between risk scores and recidivism across a 7-year period for probationers in Manitoba, Canada. The analyses defined recidivism as failure on probation for technical violations and new offenses and were based on over 14,000 offenders on probation between 1986 and 1991.<sup>60</sup> The report does not indicate whether the assaultive item was included on the scale.

More recently, Eisenberg and his colleagues examined the performance of the Wisconsin risk instrument for a sample of 42,853 Wisconsin offenders placed on community supervision in 2001-2002.<sup>61</sup> They found a correlation of  $r=.22$  between risk scores (excluding the assaultive factor) and the commitment of a new offense within three years of being placed on community supervision. Henderson and Miller examined a sample of 194 male, mostly misdemeanor offenders, released in 2000 from a Texas probation department. For the risk assessment with the assaultive item, they reported a correlation of  $r=.25$  (and an AUC of .63 for the receiver operating characteristic curve analysis) for arrest within five years of release from probation.<sup>62</sup> Latessa and his colleagues reviewed arrests for a new crime for 672 individuals on community supervision in Ohio in 2008. They found the correlation between the Wisconsin risk assessment and recidivism to be  $r=.21$ .<sup>63</sup> The researchers did not indicate whether the assaultive factor was included in the assessment, but Baird reports that the

results are actually based on the reassessment and needs instrument combined rather than the intake assessment.<sup>64</sup> The assaultive factor is not included in the reassessment instrument, and there are other differences between the two versions as well.

Several studies considered how accurately the Wisconsin risk assessment classified offenders into different risk levels as measured by subsequent recidivism. For example, revalidation studies for the Department of Corrections in Nevada and Wisconsin and for the probation departments in Orange County, California and Travis County, Texas all indicated that the recidivism rate for offenders increased with increasing classification levels of risk.<sup>65</sup> That is, offenders classified as low risk based on the Wisconsin risk scale recidivated less than offenders classified as medium risk, and both recidivated less than those classified as high risk. The Travis County revalidation included the assaultive factor (giving it a weight of 8 points) in its risk scale as did the Wisconsin revalidation (giving it a weight of 15 points). The Travis County report concluded the assaultive factor was predictive of recidivism, and the Wisconsin report concluded the factor did not adequately predict recidivism. The Nevada, Wisconsin, and Orange County reports all suggested revisions to the instrument to increase its ability to distinguish across risk levels. For example, the Orange County study indicated that a large percentage of offenders (54.8%) were classified as high risk. The study's authors suggested changing the weights for three items, eliminating one, adding a new item, and changing the cutoff scores for the classification levels. As a result

## Appendix: RNA Instrument Profile for the CAIS

---

of these changes, the percentage of offenders classified as high risk decreased to 34.5% while maintaining increasing levels of recidivism rates across the low, medium, and high classifications, and the AUC increased from .642 for the original instrument to .659 for the new instrument.<sup>66</sup>

A few studies have examined the relationship between the original Wisconsin needs assessment scale and recidivism and have found that some of the needs items are significantly related to recidivism.<sup>67</sup> However, the needs assessment scale was not specifically developed as a predictor of recidivism, and the CAIS does not report a separate needs score.

Researchers involved in the development of the CMC reported on an evaluation of the CMC in a summary article in 1986.<sup>68</sup> The evaluation followed 422 high-risk (as determined by the Wisconsin risk assessment) Milwaukee probationers randomly assigned to regular supervision, intensive supervision only, or intensive supervision as directed with CMC case planning. The study focused on three outcome measures: percentage revoked, percentage employed at termination and percentage earning income over \$400/month at termination. Although the results were in the predicted direction—the CMC with intensive supervision group performed better than the intensive supervision only group, and both performed better than the regular supervision group—only the comparison between the CMC with intensive supervision group and the regular supervision group was significant.

Researchers from the Texas Board of Pardons and Paroles followed 2,551 parolees, released during March and April 1985, for a year.<sup>69</sup> A little less than half (46%) of the parolees were supervised by parole officers trained on the CMC, and the remaining parolees served as a comparison group. All of the cases were classified as a poor, fair, or good risk based on a validated risk assessment. The CMC parolees had significantly fewer pre-revocation warrants than regular supervision parolees for the poor and fair risk groups when measured after 6- and 12-month periods. CMC parolees in the poor risk group also had significantly fewer returns to prison than non-CMC parolees. Thus CMC had the greatest effect on high risk offenders; no statistical difference was found for parolees in the good risk categories.

Researchers from the South Carolina Department of Probation, Parole and Pardon Services also found that CMC was related to outcomes for higher risk offenders—those convicted of a violent or sexual offense, who have served more than 90 days in prison, or who are under intensive supervision.<sup>70</sup> They followed two groups of offenders, matched on the basis of offense, risk score, and level of supervision, for a year during 1985-1986. One group of 200 offenders was supervised with CMC, and the other group of 219 offenders was not. The two groups differed significantly on measures of supervision failure, revocations for new offense, and revocations or unsatisfactory supervision terminations resulting in returns to prison.

CMC developers also report data from an unpublished study of 45,346 offenders in Florida placed in a community control program as an alternative to prison.<sup>71</sup>

## Appendix: RNA Instrument Profile for the CAIS

---

Approximately half of the offenders received CMC in addition to the supervision requirements for all offenders. Data for the first four years (1993 to 1997) of the program indicated that the offenders supervised with CMC had significantly lower revocation rates.

Harris and her colleagues, however, questioned the use of revocation rates as the primary indicator of success for the CMC. They suggested that officers trained on CMC techniques may be less likely to revoke offenders. They assessed the effectiveness of the CMC using three different outcome measures: write-ups for technical violations, revocations, and new arrests while under supervision. Of the 1,017 felony offenders entering probation for approximately a year beginning in March of 1991, 581 were supervised with CMC, and 436 served as the control group. CMC-supervised offenders differed significantly from offenders in the control group only on the outcome measure of revocations. In addition, the CMC group had a higher failure to comply with program conditions despite being less likely to experience revocation compared to the control group. However, an audit of the CMC-supervised cases indicated errors in implementation by probation officers, thus calling into question the extent to which CMC was implemented as intended. The authors called for more evaluations of CMC using multiple outcome measures to ensure successful revocation outcomes are due to changes in offenders' behaviors and not to officers' more tolerant supervision strategies regarding revocations for minor infractions.

**Reliability.** No information was found on the inter-rater reliability of the Wisconsin

risk instrument. Both the Wisconsin and Orange County validation studies recommended conducting inter-rater and intra-rater reliability testing to assure accurate scoring.<sup>72</sup>

The inter-rater reliability of the original needs scale was examined during its development. Probation officers listened to taped interviews with offenders and independently rated the needs of the offenders. The average rate of agreement for each of the eleven items ranged between 79% and 94% with an average overall rating of 87%.<sup>73</sup>

The report on the development of the CMC indicates that "different raters obtain the same client groups approximately 90% of the time," and agreed on individual items 70% of the time or higher, with a few exceptions.<sup>74</sup>

**Potential for bias: gender.** The revalidation of the risk instrument in Nevada; Wisconsin; Orange County, California; and Travis County, Texas all found that the instrument performed as expected for both males and females.<sup>75</sup> That is, recidivism increased across low, medium, and high categories of risk for males and females. However, as discussed under the "predictive validity" section, suggestions were made to revise the scale and cutoff scores for risk levels to improve the classification categories for all offenders.

According to the CAIS brochure, the assessment system includes "gender-specific system factors in the unique risk and needs areas of women as well as tailoring supervision strategies for women based on the most current research."<sup>76</sup> Studies comparing the recidivism rate of female

## Appendix: RNA Instrument Profile for the CAIS

---

offenders supervised based on their CAIS assessment versus those supervised without CAIS are not available in the general literature to date.

**RACE.** The revalidations of the risk instrument in Nevada; Wisconsin; and Orange County, California also found that the instrument performed as expected for Black, White, and Hispanic groups.<sup>77</sup> As with gender, suggestions were made to revise the scale and cutoff scores for risk levels to improve the classification categories for all offenders.

Studies comparing the recidivism rate of different race and ethnic groups supervised based on their CAIS assessment versus those supervised without CAIS are not available in the general literature to date.

**Independent validation.** Several of the studies cited in the validation section were conducted by independent researchers. In addition, NCCD has conducted or reported on several unpublished validation studies.

### **PRACTICAL CONSIDERATIONS.**

---

**Vendor and instrument cost.** The original Wisconsin risk and needs scales and CMC are in the public domain. CAIS is proprietary. The automated assessment and case management system is available for purchase from AutoMon and NCCD.<sup>78</sup> A subscription fee is assessed of users, the amount of which is determined on a sliding scale based on the size of the jurisdiction.<sup>79</sup> For more information, contact NCCD at [JAIS.CAIS@nccdglobal.org](mailto:JAIS.CAIS@nccdglobal.org) or AutoMon at [sales@automon.com](mailto:sales@automon.com).

**Menu of other services.** NCCD and AutoMon offer a wide array of services, training, and technical assistance to support CAIS implementation.

- **IT SERVICES.** CAIS is a web-based program available through an internet browser. The advantage of this approach is that there are no issues with infrastructure requirements and redesign of existing agency MIS systems.<sup>80</sup> AutoMon is a computer software firm that provides technology support and can customize the system to include additional assessment tools and specific reports.
- **TECHNICAL ASSISTANCE.** NCCD offers two technical assistance visits of two days each per year to CAIS clients.<sup>81</sup>
- **VALIDATION SERVICES.** NCCD will validate the risk assessment component of the CAIS for all client agencies as part of the package of services provided.<sup>82</sup> NCCD recommends conducting a revalidation study every 2-5 years, depending on the size of the jurisdiction (smaller jurisdictions may need a longer period of time to identify a large enough cohort of cases for a revalidation study). There is no added cost for this service.
- **USER TRAINING.** NCCD offers a training package that includes 24 hours of classroom work and additional follow-up practicum work.<sup>83</sup> The training is fee-based. An optional 3-day “train the trainers” course is also available and is recommended for those clients interested in developing an internally sustainable initial and refresher training program. In addition, web-based courses



## Appendix: RNA Instrument Profile for the CAIS

---

have been developed “to reduce training costs and provide greater flexibility to agencies to train new staff or provide refresher training when needed.”<sup>84</sup>

**User qualifications.** All users must take the mandatory training (see description, above) before using the CAIS.

**Administration time.** The CAIS manual reports that an assessment generally takes approximately 45 minutes to complete.<sup>85</sup>

**Modes of administration.** A semi-structured interview format is used to complete the CAIS. The CAIS manual encourages officers to follow-up on important or interesting information the offender presents during the interview.<sup>86</sup>

**Quality assurance.** When adopting any offender assessment tool, jurisdictions must be prepared to ensure appropriate implementation and proper maintenance over time. Quality assurance recommendations and guidelines for CAIS follow.

- **OVERRIDE POLICY.** The CAIS report provides an opportunity for the officer to override the risk level based on a state or local policy or at the officer’s discretion, provided a reason is given and a supervisor approves the override. The reasons for overrides vary across jurisdictions. Though some jurisdictions have made extensive use of the policy override (e.g., certain offenses automatically are placed in higher risk levels, as discussed in previous sections), discretionary overrides are less frequent. Baird reports that NCCD studies usually see overrides in the 5-7% range.<sup>87</sup>

- **FIDELITY.** CAIS offers a variety of aggregate data report options for officers and supervisors. Information regarding the implementation of the CAIS can be routinely obtained and reviewed on issues such as gender, risk levels, needs, ethnicity, worker, and unit.<sup>88</sup>
- **INSTRUMENT REVALIDATION.** Validations of the risk component are recommended every 2-5 years, depending on the size of the jurisdiction and available data.

### ENDNOTES

---

<sup>1</sup> “The goal of risk assessment is to classify offenders into different risk groups based on rates of subsequent criminal behavior” (C. Baird, personal communication, July 24, 2012). Various individuals have been involved in the development and evaluation of the different components of the CAIS; however, Christopher Baird of the National Council on Crime and Delinquency has written the most about the instrument’s development and use.

<sup>2</sup> C. Baird, personal communication, July 24, 2012.

<sup>3</sup> See p. 38 in Baird, C. (1981). Probation and parole classification: The Wisconsin model. *Corrections Today*, 43(3), 36-41. Information also provided by C. Baird, personal communication, July 24, 2012.

<sup>4</sup> See p. 12 in Baird, C. S., Heinz, R. C., & Bemus, B. J. (1979). *The Wisconsin case classification/staff deployment project: A two-year follow-up report*. Madison, WI: Wisconsin Division of Corrections.

<sup>5</sup> See pp. 8-9 in Baird, C. (2009). *A question of evidence: A critique of risk assessment models used in the justice system*. Oakland, CA: National Council on Crime and Delinquency.

<sup>6</sup> See Baird (2009) at endnote 5, pp. 9-10.

## Appendix: RNA Instrument Profile for the CAIS

---

<sup>7</sup> See pp. 2, 69 in National Council on Crime and Delinquency. (2010). *CAIS Correctional Assessment and Intervention System: System manual*. Madison, WI: Author.

<sup>8</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 8. The initial risk scale was based on an outcome measure of weighted factors that included rules violations, arrests, misdemeanor convictions, absconsions, felony convictions, and convictions for assaultive offenses. See Baird et al. (1979) at endnote 4, p. 40.

<sup>9</sup> See Baird et al. (1979) at endnote 4, p. 8.

<sup>10</sup> See Baird et al. (1979) at endnote 4, p. 8.

<sup>11</sup> See Baird et al. (1979) at endnote 4.

<sup>12</sup> Baird, C., & Neuenfeldt, D. (1990).

*Improving correctional performance through better classification: The Client Management Classification System*. Madison, WI: National Council on Crime and Delinquency.

<sup>13</sup> See National Council on Crime and Delinquency (2010) at endnote 7.

<sup>14</sup> National Council on Crime and Delinquency. (n.d.). *CAIS Correctional Assessment and Intervention System (brochure)*. Madison, WI: Author.

<sup>15</sup> See p. 154 in Harris, P. (1994). Client management classification and prediction of probation outcome. *Crime and Delinquency*, 40, 154-174.

<sup>16</sup> See p. 22 in Hubbard, D. J., Travis, L. F., & Latessa, E. J. (2001). *Case classification in community corrections: A national survey of the state of the art*. Cincinnati: University of Cincinnati. The percentages are based on the 288 agencies using a risk and needs assessment instrument. The entire sample consisted of 385 agencies.

<sup>17</sup> Taxman, F. S., Cropsey, K. L., Young, D. W., & Wexler, H. (2007). Screening, assessment, and referral practices in adult correctional settings: A national perspective. *Criminal Justice and Behavior*, 34, 1216-1234.

<sup>18</sup> W. Ore, personal communication, December 24, 2012.

<sup>19</sup> See NCCD: CAIS website at <http://www.nccdglobal.org/assessment/corr>

[ectional-assessment-and-intervention-system-cais](#).

<sup>20</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 2.

<sup>21</sup> C. Baird, personal communications, March 21 and July 24, 2012.

<sup>22</sup> See Baird et al. (1979) at endnote 4, Appendix A, pp. 39-44 for a description of the Wisconsin risk assessment instrument. Unless otherwise noted, this document is used as the source for the development description in the text.

<sup>23</sup> C. Baird, personal communication, July 24, 2012. Also see Baird et al. (1979) at endnote 4, p. 42.

<sup>24</sup> See Baird et al. (1979) at endnote 4, p. 10.

<sup>25</sup> See Baird et al. (1979) at endnote 4, p. 10.

<sup>26</sup> C. Baird, personal communications, March 21 and July 24, 2012.

<sup>27</sup> See Baird et al. (1979) at endnote 4, p. 11.

<sup>28</sup> See p. 5 in Jones, D., Johnson, S., Latessa, E., & Travis, L. (1999). Case classification in community corrections: Preliminary findings from a national survey. In *Topics in Community Corrections*, 4-8. Washington, DC: U.S. Department of Justice, National Institute of Corrections. Also see See Baird et al. (1979) at endnote 4, p. 12.

<sup>29</sup> See Baird et al. (1979) at endnote 4, p. 12.

<sup>30</sup> See Baird et al. (1979) at endnote 4, pp. 12-13.

<sup>31</sup> See Baird et al. (1979) at endnote 4, p. 12.

<sup>32</sup> See Baird et al. (1979) at endnote 4, p. 14.

<sup>33</sup> See Baird et al. (1979) at endnote 4, p. 47. Additional information provided by C. Baird, personal communication, July 24, 2012: "Some agencies use a matrix that allows them to emphasize the role of one instrument...usually the risk assessment."

<sup>34</sup> C. Baird, personal communication, July 24, 2012.

<sup>35</sup> C. Baird, personal communication, March 21, 2012.

<sup>36</sup> See Baird et al. (1979) at endnote 4, p. 18.

<sup>37</sup> See p. 257 in Lerner, K., Arling, G., & Baird, C. (1986). Client management classification strategies for case supervision. *Crime and*

## Appendix: RNA Instrument Profile for the CAIS

---

*Delinquency*, 32, 254–271. Information on the development of the CMC also is available on pp. 75–78 in National Institute of Corrections. (1981). *NIC technical assistance report: Model probation/parole management program*. Washington, DC: Author. Unless otherwise noted, the profile’s description of the CMC’s development is based on these two documents.

<sup>38</sup> The taxonomy is sometimes reported as having five supervision strategies because one strategy has a subcategory. The CAIS System Manual presents the taxonomy as five strategies. See National Council on Crime and Delinquency (2010) at endnote 7, p. 4 and p. 6. However, the CAIS demonstration report provided by the test developers presents a classification score for the original four strategies.

<sup>39</sup> The ability of an item to differentiate among the supervision strategies was tested using a chi-square analysis. See National Institute of Corrections (1981) at endnote 37, pp. 76–77.

<sup>40</sup> Items were given a rating of 1, 2, or 3. Reliability ratings for each were at least .75, .80, and .90, respectively. Chi square significance levels for an item’s ability to differentiate among supervision strategy groups were at least .05, .01, and .001, respectively. Thus an item weighted as 3 for a particular supervision strategy group had an interrater reliability of at least .9 and differentiated the supervision strategy group from the other groups at a significance level of .001 or higher. See National Institute of Corrections (1981) at endnote 37, p. 76.

<sup>41</sup> See Baird & Neuenfeldt (1990) at endnote 12. The construction and validation samples were the same: C. Baird, personal communication, July 29, 2014.

<sup>42</sup> See National Institute of Corrections (1981) at endnote 37, p. 76.

<sup>43</sup> See Lerner et al. (1986) at endnote 37, p. 258.

<sup>44</sup> Ore, W., & Baird, C. (2014, March). *Beyond risk and needs assessments*. Madison, WI: National Council on Crime and Delinquency.

<sup>45</sup> C. Baird, personal communications, March 21.

<sup>46</sup> See Ore & Baird (2014, March) at endnote 44, p. 7. Information also provided by C. Baird, personal communications, March 21.

<sup>47</sup> C. Baird, personal communications, March 21.

<sup>48</sup> See National Council on Crime and Delinquency (2010) at endnote 7, pp. 7–8.

<sup>49</sup> See National Council on Crime and Delinquency (2010) at endnote 7, pp. 7–8.

The description also is based on example CAIS demonstration reports from 2009. The reports were provided by Toni Aleman of the National Council on Crime and Delinquency, August 17, 2010.

<sup>50</sup> See National Council on Crime and Delinquency (2010) at endnote 7, pp. 45 and 65 for CAIS risk items. See Baird et al. (1979) at endnote 4, pp. 10–11 for WRN risk items.

<sup>51</sup> The table does not include the 11th item included in the original WRN instrument because it was included at the request of the Wisconsin Department of Corrections and not because of its predictive ability. The WRN items also are out of order to better compare the items across the two instruments.

<sup>52</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 3.

<sup>53</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 7.

<sup>54</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 6.

<sup>55</sup> See NCCD: CAIS webpage at <http://www.nccdglobal.org/assessment/correctional-assessment-and-intervention-system-cais>.

<sup>56</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 4.

<sup>57</sup> C. Baird, personal communication, March 21, 2012. The focus was on ensuring that certain categories “are being considered for every case by every worker, and that the

## Appendix: RNA Instrument Profile for the CAIS

---

ratings are done fairly consistently across the raters.... And so a lot of agencies either added some areas or may have deleted some areas, depending on what...input that they got from people within their agency. The needs instrument, other than looking at the inter-rater reliability, is not a research-based instrument.”

<sup>58</sup> Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34, 575-607. The effect size adjusted for sample size is  $r = .32$ . The 14 effect sizes were derived from various studies that included outcome measures of arrest, conviction, incarceration, parole violation and/or some combination. The authors referred to “risk scales” (p. 585) when describing the instruments they examined; there is no indication that effect sizes also were calculated for the Wisconsin needs scale.

<sup>59</sup> At least one of the studies included in the meta-analysis examined the predictive validity of the risk assessment instrument without the assaultive factor and reported a correlation of  $r=.17$  with recidivism. See Wright, K. N., Clear, T. R., & Dickson, P. (1984). Universal applicability of probation risk-assessment instruments. *Criminology*, 22, 113-134. As a comparison, Robinson and Porporino did include the assaultive factor in their study, giving it a weight of 15 points. They reported a correlation of  $r=.21$  between the risk score and recidivism. See Robinson, D., & Porporino, F. J. (1989, May). *Validation of an adult offender classification system for Newfoundland and Labrador*. (Research report no. R-04). Ottawa, ON: Correctional Service of Canada.

<sup>60</sup> The data are summarized in Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 18-32). Thousand Oaks, CA: Sage. The author notes that a few modifications were made to the

instrument in 1986 but does not indicate the specific changes that were made.

<sup>61</sup> Eisenberg, M., Bryl, J., & Fabelo, T. (2009, August). *Validation of the Wisconsin Department of Corrections risk assessment instrument*. New York: Council of State Governments Justice Center. The correlation decreased to .18 with the assaultive factor included on the instrument.

<sup>62</sup> Henderson, H. & Miller, H. (2013). The (twice) failure of the Wisconsin Risk Need Assessment in a sample of probationers. *Criminal Justice Policy Review*, 24, 199-221. Wisconsin test developer Baird (C. Baird, personal communication, July 24, 2012) criticized the study for using a “substantially flawed outcome measure” and a “highly selective and limited sample” as well as misrepresenting prior work on the Wisconsin Risk and Needs Assessment.

<sup>63</sup> Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). *Creation and validation of the Ohio risk assessment system: Final report*. Cincinnati, OH: University of Cincinnati Center for Criminal Justice Research.

<sup>64</sup> C. Baird, personal communication, January 27, 2012. In another communication, Baird further explained that “during the course of supervision, very high percentages of cases move to lower risk levels over time. The reclassification scale shifts emphasis from prior history items to factors that reflect behavior since the last assessment” (March 13, 2012).

<sup>65</sup> For the Nevada report, see Wagner, D., & Oremus, K. (2009, June). *Nevada Department of Public Safety Division of Parole and Probation risk and needs assessment validation*. Madison, WI: National Council on Crime and Delinquency. For the Wisconsin report, see Eisenberg et al. (2009, August) at endnote 61. For the Orange County report, see Eisenberg, M., Fabelo, T., & Tyler, J. (2011, October). *Validation of the Orange County California Probation Department risk assessment instrument*:

## Appendix: RNA Instrument Profile for the CAIS

---

*Final report.* New York: Justice Center, The Council of State Governments. For the Travis County report, see Bryl, J., Fabelo, T., & Nagy, G. (2006, August). *Travis community impact supervision: Guiding justice decisions with risk assessment instruments.*

Washington, DC: The JFA Institute.

<sup>66</sup> See Eisenberg et al. (2011, October) at endnote 65, pp. 40-46.

<sup>67</sup> For example, Robinson and Porporino (1989, May, at endnote 59) found a correlation of  $r=.14$  between the needs score and recidivism for a sample of 200 probation cases in Canada. They identified three needs items (interpersonal relationships, companions, and drug involvement) as significantly differentiating recidivists and non-recidivists (see Appendix C in report). While recidivism was related to both the risk and needs scales, neither scale differentiated well between medium and high risk offenders. This likely was due, in part, to the low base rate of recidivism (10.5%) for the entire sample. Bonta (1996, at endnote 60) found a slightly modified version of the needs assessment to be predictive for Manitoba probationers. Across a seven year period, the correlations ranged from  $r=.10$  to  $r=.22$ . In another study of over 11,000 Nevada offenders, Wagner and Oremus (2009, June, at endnote 65, pp. 27-29) found that 8 of the 11 needs items had a significant relationship to recidivism. Henderson and Miller (2013 at endnote 62) found a correlation of  $r=.19$  and an AUC=.62 between the total needs score and rearrest among a sample of 194 probationers in Texas. Their study identified three items (employment, financial management, and drug problems) as significantly related to recidivism.

<sup>68</sup> See Lerner et al. (1986) at endnote 37. The study also was reported in a publication by the National Council on Crime and Delinquency. See Ore & Baird (2014, March) at endnote 44. It is not clear from the two summaries when the data actually were collected. The Lerner et. al article says the

study was undertaken in 1979 and present data provided by the Wisconsin Division of Corrections in 1983. The Ore and Baird report refer to the “Wisconsin Study, 1986” (p. 6) and indicate that outcomes were measured 18 months after admission to probation. The size of the original sample also is unknown. Lerner et al. report that “the Ns for each outcome category varied somewhat due to missing information at termination” (p. 268). The sample size of 422 was based on those for whom information was available on revocations/discharges, the outcome measure with the most complete information.

<sup>69</sup> Eisenberg, M., & Markley, G. (1987). Something works in community supervision. *Federal Probation*, 51, 28-32.

<sup>70</sup> McManus, R. F., Stagg, D. I., & McDuffie, C. R. (1988). CMC as an effective supervision tool: The South Carolina perspective. *Perspectives, Summer*, 30-34.

<sup>71</sup> See Ore & Baird (2014, March) at endnote 44, p. 5. The study is also discussed in Harris, P. M., Gingerich, R., & Whittaker, T. A. (2004). The “effectiveness” of differential supervision. *Crime & Delinquency*, 50, 235-271.

<sup>72</sup> See Eisenberg et al. (2009, August) at endnote 61, p. 29 and Eisenberg et al. (2011, October) at endnote 65, p. 48.

<sup>73</sup> See Baird et al. (1979) at endnote 4, pp. 15-17.

<sup>74</sup> See National Institute of Corrections (1981) at endnote 37, p. 77. The report indicates that 59 CMC items had an inter-rater reliability of .9 or better, 70 items .8 or better, 97 items .7 or better, and 5 items slightly less than .7 (p. 76). Reportedly (C. Baird, personal communication, July 29, 2014), the number of items is based on the number of “forced-choice” options in the CMC and not the number of questions. Thus the number of items for which reliabilities were reported exceeds the number of questions on the instrument. At least three

## Appendix: RNA Instrument Profile for the CAIS

---

of the five raters assessed each of the 250 offenders in the sample.

<sup>75</sup> See Bryl et al. (2006, August) at endnote 65; Eisenberg et al. (2009, August) at endnote 61; Eisenberg et al. (2011, October) at endnote 65; and Wagner & Oremus (2009, July) at endnote 65. Eisenberg et al. (2009, August) also reported a correlation of  $r = -.073$ , “indicating a weak correlation between gender and new offense” (p. 23).

<sup>76</sup> See National Council on Crime and Delinquency (n.d.) at endnote 14. Also see NCCD CAIS website at <http://www.nccdglobal.org/assessment/correctional-assessment-and-intervention-system-cais>.

<sup>77</sup> See Eisenberg et al. (2009, August) at endnote 61; Eisenberg et al. (2011, October) at endnote 65; and Wagner & Oremus (2009, July) at endnote 65. Eisenberg et al. (2009, August) also reported a correlation of  $r = .05$  between race/ethnicity and new offense” (p. 25).

<sup>78</sup> See AutoMon, Assessments Management website page at <http://www.automon.com/solutions/criminal-justice/assessments>.

<sup>79</sup> C. Baird, personal communication, March 21, 2012.

<sup>80</sup> See National Council on Crime and Delinquency (n.d.) at endnote 14.

<sup>81</sup> C. Baird, personal communication, July 24, 2012.

<sup>82</sup> C. Baird, personal communication, July 24, 2012.

<sup>83</sup> C. Baird, personal communication, July 24, 2012.

<sup>84</sup> See Ore & Baird (2014, March) at endnote 44.

<sup>85</sup> See National Council on Crime and Delinquency (2010) at endnote 7.

<sup>86</sup> See National Council on Crime and Delinquency (2010) at endnote 7, p. 47.

<sup>87</sup> C. Baird, personal communication, July 24, 2012.

<sup>88</sup> See National Council on Crime and Delinquency (n.d.) at endnote 14.

# Correctional Offender Management Profile for Alternative Sanctions (COMPAS)

---

## COMPAS GLOSSARY OF TERMS

---

<b>Risk</b>	COMPAS distinguishes between risk scales, which are designed to measure the likelihood that an offender will recidivate, and needs scales, which are designed to capture information about offender needs that can be used to inform case plans and identify target criminogenic thoughts and behaviors for treatment intervention. <sup>1</sup> The authors “believe risk scales designed to predict risk should be dynamic (composed of dynamic criminogenic needs) so that one can measure changes in risk of recidivism over time.” <sup>2</sup>
<b>Static risk</b>	Authors indicate that these are “historical factors” (e.g., age at first arrest). <sup>3</sup>
<b>Dynamic risk</b>	Authors indicate that these are “criminogenic factors” (e.g., employment status, level of substance abuse). <sup>4</sup>
<b>Needs</b>	Offender needs are individual factors about the offender that, in the aggregate, have a demonstrated relationship with recidivism but that can be changed. <sup>5</sup> Included are factors such as criminal thinking, education, employment, substance abuse, residential stability and other aspects of the “person-in-environment” which guide individualized decisions in case planning. <sup>6</sup>
<b>Responsivity</b>	Responsivity refers to the principle that people respond differently to different treatment approaches. This recognizes that “the wrong treatment may make things worse and creates a need for careful matching of people to specific treatments.” <sup>7</sup> Officers who create the offender’s case plan must pay attention to responsivity issues at the intake assessment, as they capture information about the offender’s ability and readiness to make the changes to reduce their future likelihood of recidivating.
<b>Protective factors</b>	Protective factors are discussed as strengths (see below). <sup>8</sup>
<b>Strengths</b>	Offender factors (e.g., supportive families, educational and vocational strengths, stable residences in safe areas, social supports) that have shown empirical support for potential risk reduction and protecting individuals’ from the full impact of criminogenic needs. <sup>9</sup>
<b>Recidivism</b>	General recidivism refers to any new arrest within two years of the COMPAS assessment. <sup>10</sup>

---

## Appendix: RNA Instrument Profile for the COMPAS

---

### HISTORY & CURRENT USE.

---

**Creation.** COMPAS was initially developed in 1998 by the Northpointe Institute for Public Management. The instrument has since undergone several iterations of revisions and was last updated based on a national sample of 30,000 imprisoned and community-based offenders for whom COMPAS assessments were conducted between January 2004 and November 2005.<sup>11</sup> The current version of COMPAS has norms available for eight groups: male or female prison, jail, probation, or composite groups of offenders.<sup>12</sup>

**Current use.** COMPAS has been utilized by the California Department of Corrections and Rehabilitation, including probation departments in San Diego, San Francisco, Tulare, San Bernardino, and Riverside counties; Michigan Department of Corrections; New Mexico Corrections Department; New York State Department of Corrections and Community Supervision; South Carolina Department of Corrections; Wisconsin Department of Corrections; and Wyoming Department of Corrections.<sup>13</sup>

### DEVELOPMENT.

---

**Instrument purpose.** COMPAS is an automated, fourth generation risk and needs assessment instrument and case planning system.<sup>14</sup> It was “designed to help criminal justice practitioners determine the placement, supervision, and case-management of offenders in community and secure settings.”<sup>15</sup>

The COMPAS tool was designed to be adaptable for different agency decisions from

pretrial to prison release. The entire COMPAS system contains 42 separate scales that may be selected and combined for use with various offender populations (jail, prison, parole, probation) and at different decision points in the criminal justice process (pretrial release, case management).<sup>16</sup> The vendor provides client agencies with the version of the COMPAS that matches their needs.<sup>17</sup> For this reason, the actual uses and content of the COMPAS can vary substantially between agencies and between research studies. This profile focuses on the General Recidivism Risk scale and other components relevant for use with a general community-based population of adult offenders.

#### **Approach to instrument development.**

Developers of COMPAS were strongly influenced by the process used to develop an outcomes-based recidivism scale for England and Wales.<sup>18</sup> In selecting and developing risk and needs scales for the COMPAS system, Northpointe undertook a theory-guided design based upon established causal theories of crime such as low self-control theory, social learning theory, strain theory, social control theory, routine activities-opportunity theory, and a strengths and good lives perspective.<sup>19</sup> The COMPAS scales also include key offender risk and needs factors that have emerged from meta-analytic research, including the “central 8.”<sup>20</sup> All COMPAS scales are composed of items selected by instrument developers on the basis of not only their relevance to factors theoretically associated with criminal behavior but also their demonstrated statistical relationship with those constructs.<sup>21</sup>



## Appendix: RNA Instrument Profile for the COMPAS

---

COMPAS distinguishes between risk scales and needs scales.<sup>22</sup> In the development of the risk scales, researchers prioritized the use of a limited set of items (parsimony) and the ability of risk scores to predict recidivism (predictive validity).<sup>23</sup> The General Recidivism Risk scale, was statistically derived based on data from a sample of presentence investigation and probation intake cases in 2002 to predict any offense (misdemeanor or felony) arrest within two years of the offender’s COMPAS administration date.<sup>24</sup>

The needs scales (e.g., criminal thinking, education, employment, substance abuse, residential stability) capture and describe factors about the individual offender that have been found in the extant literature to be associated with criminal behavior.<sup>25</sup> These need areas are not all used in the calculation of offender recidivism risk; rather, they represent potential targets for treatment intervention to be used by the supervising officer to inform case planning efforts.<sup>26</sup>

COMPAS scales also include a mixture of both dynamic (e.g., level of substance abuse) and static items (e.g., age at first offense) to permit measurement of change over time. Although the exact items and proportion of static versus dynamic items may vary by scale and depending on the version of COMPAS used, over 50 percent of the items in COMPAS are dynamic.<sup>27</sup>

### CONTENT.

---

**Structure.** As indicated earlier, the exact structure of the COMPAS will vary by client agency. The entire COMPAS system contains 42 scales, including 4 offender recidivism

risk scales (e.g., General Recidivism Risk), 1 short 5-item recidivism risk screen scale, 19 gender-neutral “criminogenic need scales” to identify factors about the individual offender that are associated with criminal behavior in the larger population, 16 women-specific needs scales, and 2 validity scales.<sup>28</sup> The number of questions for each scale varies.<sup>29</sup>

**Items and domains.** The COMPAS Core Assessment includes 135 items that are combined into various risk and need scales.<sup>30</sup> The primary risk items within the General Recidivism Risk scale address prior criminal history, criminal associates, drug involvement, and early indicators of juvenile delinquency problems.<sup>31</sup> The 19 criminogenic need scales are organized into five overarching areas as described in Table 1.<sup>32</sup>

**Table 1. COMPAS Needs Scales**

Area	Scale
Criminal Involvement	<ul style="list-style-type: none"> <li>• Criminal Involvement</li> <li>• History of Non-Compliance</li> <li>• History of Violence</li> <li>• Current Violence</li> </ul>
Relationships/Lifestyle	<ul style="list-style-type: none"> <li>• Criminal Associates/Peers</li> <li>• Criminal Opportunity</li> <li>• Leisure/Recreation</li> <li>• Social Isolation</li> <li>• Substance Abuse</li> </ul>
Personality/Attitudes	<ul style="list-style-type: none"> <li>• Criminal Personality</li> <li>• Criminal Thinking Self Report</li> <li>• Cognitive Behavioral</li> </ul>
Family	<ul style="list-style-type: none"> <li>• Family Criminality</li> <li>• Socialization Failure</li> </ul>
Social Exclusion	<ul style="list-style-type: none"> <li>• Financial</li> <li>• Vocational/ Education</li> <li>• Social Environment</li> <li>• Residential Instability</li> <li>• Social Adjustment</li> </ul>

## Appendix: RNA Instrument Profile for the COMPAS

---

The specific items for each COMPAS scale are available from Northpointe.<sup>33</sup>

**Reporting and cutoffs.** The COMPAS software suite produces an individual assessment report to display each offender's results from the assessment tool in chart form.<sup>34</sup> Within the software application, raw scores are transformed into deciles, and each decile score is then used to determine the level of risk probability (deciles of 1-4 = low risk, 5-7 = medium risk, and 8-10 = high risk).<sup>35</sup> Cutoff scores for need scales vary by the scale with most falling into 1-5 = unlikely, 6-7 = probable, and 8-10 = highly probable.<sup>36</sup>

The decile scores and cutoffs are based upon a comparison of offender characteristics to a representative criminal population (i.e., a norming group). The norming group includes subpopulations of people from prison, jail, or probation.<sup>37</sup> Each agency has the ability to select a norming group that is most appropriate for its population of interest. For example, a probation agency might select the available probation sample as their norming group. Also, COMPAS can make use of separate norms for males and females to allow for gender-specific calibrations.

The assessment report chart of risk and needs scale results is accompanied by a narrative summary of the offender's assessment results. This document includes for each criminogenic need area a written description of the offender's need scale results, a statement from the interviewer, and a written description of associated treatment implications. Current charge and criminal history information are also presented.

COMPAS scales are also linked to specific "sets" of relevant treatment interventions and goals. These linkages are embedded within the COMPAS software and are offered as dropdown lists in the case plan section of the automated report. The lists of programs are based primarily on national evaluation research findings and the broader research literature with an emphasis on cognitive behavioral interventions, while simultaneously excluding programs shown to be ineffective by current evaluation research.<sup>38</sup> Program lists can be modified by client users based upon local knowledge of program effectiveness.

### INSTRUMENT RELIABILITY AND VALIDITY.

---

**Populations studied.** A number of internal and external validation studies have been conducted on COMPAS. These studies have focused on the use of the tool by the Michigan Department of Corrections,<sup>39</sup> New York State Division of Parole,<sup>40</sup> New York State Division of Probation and Correctional Alternatives,<sup>41</sup> and California Department of Corrections and Rehabilitation.<sup>42</sup> When implementing COMPAS in a new jurisdiction, Northpointe researchers typically incorporate an outcomes study with at least a year of follow-up for an initial analysis.<sup>43</sup>

Brennan, Dieterich and Ehret report that the COMPAS General Recidivism Risk scale also has been validated internally by Northpointe using "multi-year prospective outcome studies in new samples as well as for different racial/ethnic and gender groups across different state systems."<sup>44</sup> However, no comprehensive research publication of

## Appendix: RNA Instrument Profile for the COMPAS

---

these studies is publicly available at this time.

**Predictive validity.** The predictive validity of the COMPAS General Recidivism Risk scale has been examined in multiple internal pilot tests and outcome studies. Test developers report predictive validity Area Under the Curve (AUC) values ranging from .66 to .73 for any arrest.<sup>45</sup> An independent validation conducted on a sample of California parolees by Farabee and his colleagues reported an AUC of .70 for predicting any arrest within two years of being released from prison.<sup>46</sup> A study of 57 New York state probation departments using the COMPAS reported an AUC of .71 for predicting rearrest within two years among a sample of offenders admitted to probation in 2009.<sup>47</sup>

The Northpointe Practitioners Guide to COMPAS also reports on two studies examining the predictive validity of the COMPAS needs scales.<sup>48</sup> The first reports correlations ranging from  $r = -.07$  to  $r = .28$  and AUC values ranging from .51 to .63 across the 19 scales. The second study reports correlations ranging from  $r = -.16$  to  $r = .27$  and AUC values ranging from .50 to .66 across 18 of the scales.

**Reliability.** The test developers report average alpha scores measuring internal consistency of  $r = .70$  in a study of California prisoners and  $r = .73$  in a study of San Bernardino probationers.<sup>49</sup> They also report the alpha values for a combined sample of 47,679 males from California and Michigan Departments of Corrections ranging from  $r = .53$  to  $r = .86$ .<sup>50</sup>

With regard to test-retest reliability, Farabee and his colleagues reported correlations for COMPAS scales that ranged from .7 to 1.00 with an overall average score of .88, indicating that different assessment administrators provide consistent scoring of scale items.<sup>51</sup>

**Potential for bias.** The test developers report that they excluded all items that had any mention of racial, gender, religious or national origin issues in the assessment. They also report that the COMPAS scales show no systematic differences by race and gender on tests of internal consistency (Cronbach's alpha).<sup>52</sup>

- **GENDER.** COMPAS has gender-specific norm groups—female offender scores are compared to the scores of other females. Test developer Brennan and his colleagues report the predictive validity of COMPAS results did not differ significantly between men and women.<sup>53</sup> The test developers also report that COMPAS also now includes the new gender-responsive assessment designed and validated by Van Voorhis and colleagues at the University of Cincinnati.<sup>54</sup>
- **RACE.** COMPAS developers report finding very little variation in predictive validity between racial/ethnic groups.<sup>55</sup> One independent study came to a different conclusion about the predictive validity of the tool with minority offenders, concluding that the tool is only valid for use with Caucasians.<sup>56</sup> Northpointe researchers, however, argue that the sample size and base rates of

## Appendix: RNA Instrument Profile for the COMPAS

---

offending in the study were insufficient to address the question.<sup>57</sup>

**Independent validation.** A few independent evaluations of COMPAS have been conducted with mixed findings. Farabee and colleagues examined 91,334 parolees in California who had been assessed with COMPAS prior to release and had been on parole for at least one year. They concluded that the COMPAS had high test-retest reliability and acceptable predictive validity for the general recidivism risk scale.<sup>58</sup> As noted above, Fass and colleagues examined the predictive validity of the COMPAS using a male cohort of offenders released into the community from New Jersey prisons between 1999 and 2002, with a post-release outcome period of twelve months and found the COMPAS most predictive of Caucasian recidivism and least predictive of African American recidivism.<sup>59</sup> A third review by Skeem and Loudon examined the COMPAS based upon a synthesis of three extant reports.<sup>60</sup> They concluded that the COMPAS is relatively easy for professionals to apply and has internal consistency reliability. The authors concluded that there was no sound evidence to indicate predictive validity, construct/content validity, or high inter-rater reliability of the COMPAS.<sup>61</sup> Northpointe researchers contend that Skeem and Loudon's conclusions are invalid because their review was based on ongoing outcome studies with preliminary and incomplete data.<sup>62</sup>

### **PRACTICAL CONSIDERATIONS.**

---

**Vendor and instrument cost.** The COMPAS is a proprietary tool offered by

Northpointe, Inc. For more information on the instrument and software packages available, refer to their website at [www.northpointeinc.com](http://www.northpointeinc.com).<sup>63</sup>

**Menu of other services.** COMPAS offers a wide array of services, training, and technical assistance to support implementation.<sup>64</sup>

- **IT SERVICES.** The COMPAS software is scalable depending on client decision support needs and includes a user-configurable case planning module that is prepopulated with offender needs assessment results. IT customization services are available from Northpointe, and clients also can opt to have the application hosted on Northpointe's system rather than integrated into the client's system.<sup>65</sup>
- **VALIDATION SERVICES.** Northpointe offers clients an array of research services, including local validation research studies. The costs will vary depending on sample size, length of outcome, and the scope of the study (e.g., overall predictive validity or breakdowns by gender, race, ethnicity, and/or other factors).<sup>66</sup>
- **REASSESSMENTS.** Offender reassessment is built into the software to allow direct comparisons of offender profiles across time.<sup>67</sup> Northpointe leaves the decision to reassess to the discretion of the agency based on factors such as case management goals and objectives, length of time the offender is under supervision, staff resources, and so forth. If an agency opts to reassess, Northpointe suggests that it be conducted at least 8 to 12 months after the initial COMPAS

## Appendix: RNA Instrument Profile for the COMPAS

---

assessment to better measure true offender changes.<sup>68</sup>

- **USER TRAINING.** The COMPAS standard two-day training is mandatory and is typically included in the purchase of the system.<sup>69</sup> Users are instructed on how to use the tool, interpret assessment results, and create case plans for offenders to address high-need areas.<sup>70</sup> Northpointe also offers additional training options, depending on an agency's needs.<sup>71</sup>
- **CUSTOMER SUPPORT.** Northpointe provides customer and technical support Monday through Friday, 8AM to 5PM ET.<sup>72</sup>

**User qualifications.** Any user of the tool must complete a two-day COMPAS user training.<sup>73</sup> The instrument can be used by those with limited computer experience and education.<sup>74</sup>

**Administration time.** Depending on the version of COMPAS selected for use by the client agency, the assessment may take between ten minutes to one-hour.<sup>75</sup>

In an independent survey study of Parolee Services administrators in California, test administrators reported taking an average of 39 minutes to administer the COMPAS re-entry assessment interview, 58 minutes reviewing an offender's file, and 24 minutes to enter the results into the database.<sup>76</sup>

**Modes of administration.** COMPAS relies upon three procedures to collect information. First, data are gathered from official records by a criminal justice professional. Second, a trained test administrator conducts a structured

interview with the offender. Third, offenders complete a self-reported paper and pencil questionnaire. Each data modality accounts for about one-third of the data collected.<sup>77</sup>

The entire COMPAS system is automated, but requires manual input of the raw data collected by the test administrator. In some instances the official criminal records can automatically populate the criminal history section of COMPAS, where the appropriate transfer software is present.

Northpointe also offers an Ad-Hoc Report Generator that allows for client customization of various management and monitoring reports. These reports can be exported into PDF format or excel, word, XML, or RFT for import into statistical packages for further analysis.<sup>78</sup>

**Quality assurance.** When adopting any offender assessment tool, jurisdictions must be prepared to ensure appropriate implementation and proper maintenance over time. Quality assurance recommendations and guidelines for the COMPAS follow.

- **OVERRIDE POLICY.** COMPAS designers expect staff to disagree with COMPAS in about ten percent of cases due to mitigating or aggravating circumstances.<sup>79</sup> Northpointe defines mitigating factors as those that "may excuse the offender, reduce the seriousness of the crime or raise the likelihood of a pro-social adjustment."<sup>80</sup> They define aggravating factors as "extraneous information that makes the offense more serious, more violent, or may appear to make the offender more culpable, more resistant to treatment,

## Appendix: RNA Instrument Profile for the COMPAS

---

and so forth.”<sup>81</sup> In these cases, staff is encouraged to use their professional judgment to override the scale results. It is suggested that staff document the override reason and make such reasons available to supervisory staff for monitoring.<sup>82</sup>

- **FIDELITY.** COMPAS’ multi-modal data collection is designed to promote assessment reliability and ensure corroboration of offender responses. In addition, COMPAS Core contains two scales designed to examine the validity of offender responses to self-report items. One of these scales tests the offender for extreme responses (the Lie Scale) and the other tests offender responses for consistency (the Random Responding Scale).<sup>83</sup> These scales were introduced as a means to detect when offenders deliberately provide false responses to self-report items, signaling to the test administrator that further scrutiny may be required.
- **INSTRUMENT REVALIDATION.** Northpointe encourages periodic local validation of the COMPAS, as frequently as every other year.<sup>84</sup> However, they note that they have not yet found any “statistically significant deviations” in local validations of the COMPAS from national norm group studies.<sup>85</sup>

### ENDNOTES

---

<sup>1</sup> Northpointe Institute for Public Management (2012). *Practitioner’s guide to COMPAS*. Traverse City, MI: Author.

<sup>2</sup> Northpointe (2012) at endnote 1, p. 15.

<sup>3</sup> See Northpointe (2012) at endnote 1, p. 1.

<sup>4</sup> See Northpointe (2012) at endnote 1, p. 1.

<sup>5</sup> See Northpointe (2012) at endnote 1, p. 17.

<sup>6</sup> See Northpointe (2012) at endnote 1, p. 17.

<sup>7</sup> See Northpointe (2012) at endnote 1, p. 44.

<sup>8</sup> See p. 23 in Brennan, T. Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36, 21-40.

<sup>9</sup> See Brennan et al. (2009) at endnote 8, p. 23. Information also provided by T. Brennan, personal communication, August 15, 2012.

<sup>10</sup> See p. 15 in Northpointe Institute for Public Management. (2010). *Practitioner’s guide to COMPAS*. Traverse City, MI: Author.

<sup>11</sup> See Northpointe (2012) at endnote 1, p. 2.

<sup>12</sup> See Northpointe (2012) at endnote 1, p. 9.

<sup>13</sup> This list may not be exhaustive of all locations that have used or are using COMPAS. The list is drawn from Brennan, T., Dieterich, W., & Ehret, B. (2007). *Research synthesis: Reliability and validity of COMPAS*. Traverse City, MI: Northpointe; Northpointe (2012) at endnote 1, p. 15; the Wisconsin Department of Corrections website at <http://doc.wi.gov/about/doc-overview/office-of-the-secretary/office-of-reentry/compas-assessment-tool>; and W. Dieterich, personal communication, June 16, 2014.

<sup>14</sup> See Northpointe (2012) at endnote 1, p. 1.

<sup>15</sup> See p. 3 in Farabee, D., S. Zhang, R. Roberts, and J. Yang (2010). *COMPAS validation study: Final report* (Tech. Rep.). Los Angeles: Semel Institute for Neuroscience and Human Behavior University of California.

<sup>16</sup> See Northpointe (2012) at endnote 1, p. 2.

<sup>17</sup> See Northpointe (2012) at endnote 1, p. 2. For example, the specialized REENTRY COMPAS was designed for use with longer-term prison inmates, who may have distinct needs following a lengthy period of incarceration (e.g., social support, housing needs). T. Brennan, personal communication, August 15, 2012.

<sup>18</sup> See Northpointe (2012) at endnote 1, p. 15.

## Appendix: RNA Instrument Profile for the COMPAS

---

<sup>19</sup> See Northpointe (2012) at endnote 1, pp. 6-8 and Brennan, et al. (2009) at endnote 8, p. 3.

<sup>20</sup> See Brennan, et al. (2009) at endnote 8, p. 3. The central (or big) eight factors include: antisocial attitudes, antisocial associates, a history of antisocial behavior, antisocial personality pattern, problematic circumstances at home, difficulties at work or school, problems with leisure activities, and substance abuse. See Andrews, D., J. Bonta, and S. Wormith (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency*, 7-27.

<sup>21</sup> See Northpointe (2012) at endnote 1, p. 15, 17. Also see p. 16 in Brennan, T., Dieterich, W., Breitenbach, M., & Mattson, B. (2009). *A response to "Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)."*  Traverse City, MI: Author.

<sup>22</sup> See Northpointe (2012) at endnote 1, p. 14.

<sup>23</sup> See Northpointe (2010) at endnote 10, p. 15.

<sup>24</sup> See Northpointe (2010) at endnote 10, pp. 15, 24. The risk scale was derived from a regression model with bootstrap validation. Though not the subject of this profile, the other COMPAS risk scales (Violent Recidivism Risk, Pretrial Release Risk or Failure-to-Appear Risk, and Community Non-Compliance Risk) were similarly developed using different construction samples of offenders. The Violent Recidivism Risk Scale was developed in 2006 based on a sample of presentence investigation and probation intake cases to predict violent offenses within two years of intake. The Pretrial Release Risk Scale was developed in 2010 based upon felony defendants assessed with COMPAS in Kent County, Michigan. The scale is used to predict failure to appear and new felony arrests. See Northpointe (2010) at endnote 10, pp. 15-16. The Community Non-Compliance Risk Scale was designed to predict community non-compliance (likelihood of offender technical

violations) for probation and parole agencies. It is currently undergoing revision by the instrument developers. T. Brennan, personal communication, August 15, 2012.

<sup>25</sup> See Northpointe (2012) at endnote 1, p. 17.

<sup>26</sup> Research developers applied psychometric procedures such as factor analysis to maximize scale coherence, or logical consistency; see Brennan, T., Dieterich, W., & Ehret, B. (2007). *Research synthesis: Reliability and validity of COMPAS*. Traverse City, MI: Northpointe Institute for Public Management.

<sup>27</sup> T. Brennan, personal communication, August 15, 2012.

<sup>28</sup> See Northpointe (2012) at endnote 1, p. 2. Though not discussed in the *Practitioners Guide to COMPAS*, Community Non-compliance scale, currently under revision, is one of the four risk scales; T. Brennan, personal communication, August 15, 2012. The 5-item recidivism risk screen scale is a new addition to the system; W. Dieterich, personal communication, June 16, 2014.

<sup>29</sup> Northpointe Institute for Public Management. (2008). *Measurement & treatment implications of adult COMPAS scales*. Traverse City, MI: Author.

<sup>30</sup> L. Morris, personal communication, August 11, 2010.

<sup>31</sup> See Northpointe (2012) at endnote 1, p. 24.

<sup>32</sup> L. Morris, personal communication, August 11, 2010.

<sup>33</sup> See Northpointe (2008) at endnote 29.

<sup>34</sup> See example in COMPAS flyer: Northpointe Institute for Public Management. (n.d.). *COMPAS/core*. Williamsburg, MI: Author (available [http://www.northpointeinc.com/files/downloads/COMPAS\\_Core\\_Flyer\\_Front.pdf](http://www.northpointeinc.com/files/downloads/COMPAS_Core_Flyer_Front.pdf) and [http://www.northpointeinc.com/files/downloads/COMPAS\\_Core\\_Flyer\\_Back.pdf](http://www.northpointeinc.com/files/downloads/COMPAS_Core_Flyer_Back.pdf).)

<sup>35</sup> See Northpointe (2012) at endnote 1, pp. 9-11. Decile scores are provided as a percentile ranking system that compares the offender's raw score on a particular scale with the scores from a normative group. For example,

## Appendix: RNA Instrument Profile for the COMPAS

---

a decile score of “1” on a particular scale indicates that the offender’s score falls in the lowest 10% of all offenders in the normative group.

<sup>36</sup> See Northpointe (2012) at endnote 1, p. 11.

<sup>37</sup> See Northpointe (2012) at endnote 1, p. 5.

<sup>38</sup> T. Brennan, personal communication, August 15, 2012.

<sup>39</sup> Brennan, T. & Dieterich, W. (2008). *Michigan Department of Corrections Core COMPAS pilot study: One-year follow-up* (Tech. Rep.). Traverse City, MI: Northpointe. Dieterich, W., Brennan, T. & Oliver, W. (2011). *Predictive validity of the Reentry COMPAS Core risk scales: A probation outcomes study conducted for the Michigan Department of Correction* (Tech Rep.). Traverse City, MI: Northpointe.

<sup>40</sup> Brennan, T., Dieterich, W., & Breitenbach, M. (2008). *New York State Division of Parole COMPAS Reentry pilot study: Two-year follow-up: Updated predictive models*. (Tech Rep.). Traverse City, MI: Northpointe.

<sup>41</sup> Brennan, T., & Dieterich, W. (2009). *Testing the predictive validity of the DPCA COMPAS risk scales: Phase I* (Tech. Rep.). Traverse City, MI: Northpointe.

<sup>42</sup> See Farabee et al. (2010) at endnote 15. Also see review by Skeem, J., & Louden, J. E. (2007). *Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*. Davis, CA: Center for Public Policy Research.

<sup>43</sup> See Brennan et al. (2009) at endnote 21, p. 8.

<sup>44</sup> See Brennan et al. (2009) at endnote 21, p. 8.

<sup>45</sup> See Brennan et al. (2009) at endnote 8, p. 30 and Northpointe (2012) at endnote 1, p. 16. The AUC’s reported for the New York probation study (n=2,328) differed in the two references (.66 for the former and .71 for the latter) because of different analyses: W. Dieterich, personal communication, June 7, 2014.

<sup>46</sup> See Farabee (2010) at endnote 15, p. 24.

The AUC value for predicting a rearrest for a violent offense within two years was .65.

<sup>47</sup> Lansing, S. (2012, September). *New York State COMPAS-probation risk and need assessment study: Examining the recidivism scale’s effectiveness and predictive accuracy*. Albany, NY: Division of Criminal Justice Services, Office of Justice Research and Performance.

<sup>48</sup> Northpointe Institute for Public Management (2013). *Practitioner’s guide to COMPAS*. Traverse City, MI: Author. The authors report that the data used to compute the predictive validity of the needs scales come from the study by Farabee et al. (2010) at endnote 15 and Brennan et al. (2009) at endnote 8. The predictive validity scales were not reported in the original articles. The Farabee et al. study included 23,635 offenders who were followed for two years after their release from prison. The outcome measure was any arrest. The Brennan et al. study included 2,328 probationers who were followed for one year after intake. The outcome measure was felony arrest.

<sup>49</sup> See Northpointe (2012) at endnote 1, p. 21.

<sup>50</sup> See Northpointe (2012) at endnote 1, p. 22.

<sup>51</sup> Reliability scores are based on all scales without a distinction being made between the General Recidivism and the Violent Recidivism scales. See Farabee et al. (2010) at endnote 15, pp. 13-14.

<sup>52</sup> T. Brennan, personal communication, August 15, 2012.

<sup>53</sup> See Brennan et al. (2009) at endnote 8, p. 33.

<sup>54</sup> See, for example, Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women’s risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior*, 37, 261-288. Inclusion in COMPAS noted in T. Brennan, personal communication, August 15, 2012.

<sup>55</sup> See Brennan et al. (2009) at endnote 8, p. 33.



## Appendix: RNA Instrument Profile for the COMPAS

---

<sup>56</sup> Fass, A. W., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice and Behavior*, 35, 1095-1108. Fass and colleagues examined the predictive validity of the COMPAS for different racial and ethnic groups and reported an AUC value of .81 for male Caucasians, .48 for male African Americans, and .67 for male Hispanics.

<sup>57</sup> See Brennan et al. (2009) at endnote 8, p. 33. Also see Northpointe (2012) at endnote 1, p. 16.

<sup>58</sup> See Farabee et al. (2010) at endnote 15, pp. 3-4.

<sup>59</sup> See Fass et al. (2008) at endnote 56, p. 7.

<sup>60</sup> See Skeem, J., & Loudon, J. E. (2007) at endnote 42.

<sup>61</sup> See Skeem, J., & Loudon, J. E. (2007) at endnote 42, p. 6.

<sup>62</sup> See Brennan et al. (2009) at endnote 21.

<sup>63</sup> Additional contact is available with Northpointe representatives by phone at 888-221-4615 or by email at [info@npipm.com](mailto:info@npipm.com).

<sup>64</sup> See Northpointe (n.d.) at endnote 34.

<sup>65</sup> See Northpointe's Technology web page at <http://www.northpointeinc.com/services/technology>. Also see pp. 9-11 in Northpointe Institute for Public Management (2012). *COMPAS risk & need assessment system: Selected questions posed by inquiring agencies*. Traverse City, MI: Author (available [http://www.northpointeinc.com/files/downloads/FAQ\\_Document.pdf](http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf)).

<sup>66</sup> See Northpointe Institute for Public Management (2012) at endnote 65, pp. 6, 9. Information also provided by T. Brennan, personal communication, August 15, 2012.

<sup>67</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 8. Information also provided by T. Brennan, personal communication, August 15, 2012.

<sup>68</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 8.

<sup>69</sup> T. Brennan, personal communication, August 15, 2012.

<sup>70</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 7.

<sup>71</sup> See Northpointe's Training web page at <http://www.northpointeinc.com/services/training> and Northpointe (2012) at endnote 65, pp. 7, 9.

<sup>72</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 9.

<sup>73</sup> T. Brennan, personal communication, August 15, 2012.

<sup>74</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 7.

<sup>75</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 1.

<sup>76</sup> See Farabee et al. (2010) at endnote 15.

<sup>77</sup> See Brennan et al. (2009) at endnote 21, pp. 21-22. Information also from T. Brennan, personal communication, August 15, 2012.

<sup>78</sup> See Northpointe's Ad Hoc Report Generator web page at <http://www.northpointeinc.com/products/ad-hoc-report-generator> and Northpointe (2012) at endnote 65, pp. 8, 10.

<sup>79</sup> See Northpointe (2012) at endnote 1, p. 28.

<sup>80</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 6.

<sup>81</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 6.

<sup>82</sup> See Northpointe Institute for Public Management (2012) at endnote 1, p. 28.

<sup>83</sup> See Northpointe (2012) at endnote 1, pp. 42-43.

<sup>84</sup> T. Brennan, personal communication, August 15, 2012.

<sup>85</sup> See Northpointe Institute for Public Management (2012) at endnote 65, p. 6.

# Level of Service Assessments: Level of Service Inventory-Revised (LSI-R) Level of Service/Case Management Inventory (LS/CMI)

---

## LS GLOSSARY OF TERMS

---

<b>Risk</b>	Risk factors refer “to characteristics of people and their circumstances that are associated with an increased chance of future criminal activity.” <sup>1</sup> An offender’s risk level is important to decisions of release, supervision, and the allocation of treatment resources. “According to the risk principle of case classification, more intensive services are best allocated to the higher-risk cases while low-risk cases have a low probability of recidivism even in the absence of treatment services.” <sup>2</sup> Level of risk is measured with both static and dynamic risk factors (see below).
<b>Static risk</b>	Static risk factors are fixed or stable offender characteristics and aspects of personal history, such as age of first offense, that are related to the risk of reoffending. <sup>3</sup>
<b>Dynamic risk</b>	Dynamic risk factors, also referred to as criminogenic need factors, refer to risk factors that can change (e.g., antisocial attitudes) and thus “suggest appropriate intermediate targets” for reducing recidivism. <sup>4</sup>
<b>Needs</b>	Authors differentiate between criminogenic needs – problematic circumstances related to the risk of reoffending (see dynamic risk, above) – and noncriminogenic needs – problematic circumstances (e.g., homelessness) not related to the risk of reoffending. <sup>5</sup>
<b>Responsivity</b>	Responsivity refers to delivering treatment programs consistent with an offender’s ability and learning style. General responsivity – using social learning and cognitive-behavioral principles to change behavior – is distinguished from specific responsivity – offender characteristics (e.g., cognitive development) that may affect an offender’s success in a program. <sup>6</sup> Responsivity characteristics are not necessarily related to risk, “but they should be considered, particularly in the planning of intervention strategies.” <sup>7</sup>
<b>Protective factors</b>	The authors use the terms “strength” (see below) and “protective” factors synonymously. <sup>8</sup>
<b>Strengths</b>	“Strengths refer to characteristics of people and their circumstances that are associated with reduced chances of criminal activity.” <sup>9</sup> Strengths “may serve as protective factors and actively reduce the chances of antisocial conduct.” <sup>10</sup>
<b>Recidivism</b>	Recidivism has been defined variously (e.g., new arrest, new conviction, new incarceration) across studies examining the LS instruments. <sup>11</sup>

---

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

### HISTORY & CURRENT USE

---

**Creation.** In the late 1970s, Canadian psychologist Don Andrews consulted with the Ontario Ministry of Correctional Services to develop a convenient, standardized, and reasonably comprehensive tool to record offender attributes and help probation officers make decisions about the level of supervision an offender would need.<sup>12</sup> Initial versions of the tool were tested and refined by Andrews and colleague James Bonta and subsequently published as *The Level of Service Inventory-Revised (LSI-R)* in 1995.<sup>13</sup> This instrument is still available and widely used; however, in 2004, Andrews and Bonta, joined by colleague J. Stephen Wormith, published an updated version of the LSI-R, the *Level of Service/Case Management Inventory (LS/CMI)*, that includes additional sections designed to help with generating an offender's case plan and monitoring progress on its implementation.<sup>14</sup>

**Current use.** As of 2010, the LS instruments' developers report widespread use of the assessments, including jurisdictions in 23 states and Puerto Rico in America, 9 Canadian jurisdictions, and several other countries around the world.<sup>15</sup>

### DEVELOPMENT

---

**Instrument purpose.** The LS instruments are designed to help probation, parole, and other correctional officers identify areas of risk and needs that can be addressed with programming to reduce offender risk while applying the least restrictive and onerous supervision necessary for safety.<sup>16</sup> The instrument developers sought to make the

information used for risk and treatment decisions transparent and consistent across correctional officers.<sup>17</sup> In addition, the LS/CMI focuses on additional information relevant to case management, treatment planning, and service delivery.<sup>18</sup>

**Approach to instrument development.** Personality and social learning perspectives of criminal conduct, research on recidivism, and the professional opinions of probation officers guided the development of the LS instruments.<sup>19</sup> From these sources of information, a large list of potential items was generated and subsequently screened for redundancy, theoretical consistency, and predictive ability.<sup>20</sup> Through this process, the instrument's developers identified a set of items they thought predictive of recidivism *and* useful for case management and treatment planning. The latter purpose they considered crucial for helping probation officers identify intermediate targets of change. Thus the developers took a more comprehensive approach to item selection rather than identifying the minimum number of items most predictive of recidivism alone.<sup>21</sup>

Ottawa probation officers began scoring offenders on the LSI-VI version of the instrument, consisting of 58 items, in the summer of 1980. The first 598 offenders receiving the LSI-VI assessment served as the initial validation sample. This early evaluation demonstrated that LSI-VI scores were related to probation officers' risk decisions and in-program recidivism outcomes.<sup>22</sup> Following additional testing and refinement, the developers eventually published the 54-item LSI-Revised in 1995.<sup>23</sup>

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

Canadian samples of 956 male offenders from two detention centers and a jail, and 1,414 female offenders from a medium security institution for adult women, serve as the norm (or reference) groups for assessing an offender's risk level.<sup>24</sup> Normative data from the U. S., added in 2003, consists of 23,721 male and females who are in community corrections or incarcerated.<sup>25</sup> The community offender sample—those on probation or parole—includes 4,240 individuals from seven samples in two midwestern states, one southern state, and one northeastern state.<sup>26</sup> Normative data are provided for male inmates, male community offenders, female inmates and female community offenders.

Beginning in 1994, the Ontario Ministry of Community Safety and Correctional Services initiated a review of the LSI-R to address users' concerns (e.g., validity of the LSI-R with specific types of offenders, the omission of strengths and noncriminogenic needs) and produce training materials that better linked LSI-R use with evidence-based correctional practices.<sup>27</sup> The review involved broad consultation with representatives of community and institutional corrections, research and training units, and a variety of related government offices and professional associations. This feedback led to the development of the LSI-Ontario Revision which was the foundation for the LS/CMI. In particular, the LS/CMI's manual and scoring instructions were modified for application to a wider range of jurisdictions.<sup>28</sup>

Differences between the LSI-R and the LS/CMI are the latter's greater focus on the central 8 factors identified in the research literature as most predictive of recidivism

and the elimination of items with no or very low correlation with recidivism in calculating the risk/need level.<sup>29</sup> In addition, new sections were added to the LS/CMI to sample case strengths, responsivity considerations, specific risk/need factors and noncriminogenic needs.<sup>30</sup>

The LS/CMI was developed based on the results of studies conducted by the Ontario Ministry of Community Safety and Correctional Services. The LS/CMI offers Canadian, U.S., United Kingdom, and Singaporean normative data as well as data on young offenders.<sup>31</sup> The U. S. normative data is based on 48,384 offenders from nine geographically diverse jurisdictions.<sup>32</sup> Like the LSI-R, norms are available for four groups: male inmates, male community offenders, female inmates and female community offenders.

### CONTENT

---

**Structure.** Both the LSI-R and the LS/CMI calculate a single risk and needs score. Items are scored or recoded as either yes (1) or no (0) and then summed for the total score. Both instruments include a profile form that easily converts the raw score to a percentile, based on the appropriate normative group.

The LS/CMI includes ten additional sections that gather data on factors that may influence an offender's behavior. These include:

- Specific Risk/Need Factors (Section 2),
- Prison Experience—Institutional Factors (Section 3),
- Other Client Issues (Section 4),

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

- Special Responsivity Considerations (Section 5),
- Risk/Need Summary and Override (Section 6),
- Risk/Need Profile (Section 7),
- Program/Placement Decision (Section 8),
- Case Management Plan (Section 9),
- Progress Record (Section 10), and
- Discharge Summary (Section 11).<sup>33</sup>

These additional sections are not scored; they provide qualitative information important to supervision and treatment decisions.

**Items and domains.** The LSI-R consists of 54 items across 10 subcomponents, and the LS/CMI consists of 43 items across 8 subcomponents (see Table 1).<sup>34</sup> Both instruments include static and dynamic risk items.

**Table 1. LSI-R and LS/CMI Subcomponents**

Subcomponent	# of Items	
	LSI-R	LS/CMI
Criminal History	10	8
Education/Employment	10	9
Financial	2	
Family/Marital	4	4
Accommodation	3	
Leisure/Recreation	2	2
Companions	5	4
Alcohol/Drug Problems	9	8
Emotional/Personal	5	
Attitudes/Orientation	4	4*
Antisocial Pattern		4

\*Renamed Procriminal Attitude/Orientation in LS/CMI

The LS/CMI omits the financial and accommodation subcomponents of the LSI-

R. In addition, the LS/CMI has a new subcomponent called antisocial pattern which is comprised of some of the emotional/personal items on the LSI-R. The LS/CMI also allows the test administrator to indicate whether a subcomponent is considered a strength for the offender and thus could be used in case planning to help address other problem areas.

**Reporting risk levels.** The LSI-R groups offenders on probation into three levels of risk (minimum, medium, maximum) based on their overall score.<sup>35</sup> The LS/CMI groups offenders on probation into 5 levels of risk (very low, low, medium, high, very high) based on their overall scores.<sup>36</sup> The cutoff scores are the same for males and females.

The test developers suggest a range of total risk and needs scores to include in each risk level; however, they strongly recommend that jurisdictions develop their own classifications based on research and local considerations such as staff resources, tolerance for failure, and available security options.

Another step in interpreting the results of the assessment is to consider the offender's score on each subcomponent. Those subcomponents with higher scores indicate areas to address in the offender's case plan.<sup>37</sup> The LS/CMI includes "Section 7: Risk/Need Profile" in which each subcomponent score is transferred to a table that identifies the risk/need level for that subcomponent. Thus the offender is rated very low, low, medium, high, or very high on each subcomponent, too.<sup>38</sup> In addition, the LS/CMI suggests that any subcomponent designated a "strength" also should be considered in developing an offender's case plan.

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

### INSTRUMENT RELIABILITY AND VALIDITY.

---

**Populations studied.** As noted under the “Development” section, the normative groups for the LS instruments include males and females, adults and juveniles, individuals in a range of correctional settings, and individuals from several different countries. In addition, other researchers have studied the instruments in a variety of jurisdictions; some examples of these studies follow.

**Predictive validity.** A 2013 meta-analytic review of 30 years of research on the LS scales conducted by Olver and his colleagues found that, across 124 samples and a total of 130,833 adult and juvenile offenders from around the world, the LS total scores significantly predicted general community recidivism ( $r_w = .30$  and  $.29$  for fixed- and random-effects models, respectively).<sup>39</sup> Vose and her colleagues’ 2008 review of 47 studies involving adults, juveniles, males and females in a variety of correctional placements in the United States, Canada, and Europe found a statistically significant relationship between total LS score and recidivism in 81% of the studies and a positive relationship between LS scores and recidivism in 98% of the studies.<sup>40</sup> The correlations across studies examining new charges, re-arrest, reconviction, and reincarceration ranged from  $r = .06$  to  $r = .51$ .<sup>41</sup> A 1996 meta-analysis by Gendreau and his colleagues yielded a mean effect size of  $r = .35$ .<sup>42</sup> A second meta-analysis by Gendreau and his colleagues in 2002 resulted in a mean effect size of  $r = .37$ .<sup>43</sup> Based on these meta-analyses and an additional study by Hemphill and Hare; Andrews, Bonta, and

Wormith summarized the effect size of the LSI-R in a 2006 article as  $.36$  for predicting general recidivism.<sup>44</sup> In addition, studies using receiver operating characteristic analysis have reported areas under the curve (AUC) of  $.689$  for a sample of federal probationers,  $.644$  for a sample of Iowa probationers, and  $.652$  for a sample of Iowa parolees.<sup>45</sup>

It should be noted that most of these studies have focused on the predictive validity of the LSI-R. Because Section 1 of the LS/CMI is highly correlated with the LSI-R, the test developers believe the predictive validity of the LS/CMI is equal or better than the LSI-R.<sup>46</sup> At least one study confirms their belief, finding a correlation of  $r = .39$  between LS/CMI total risk scores and recidivism.<sup>47</sup>

**Dynamic predictive validity.** A few studies have examined whether changes in LSI-R scores over time are related to changes in recidivism rates. Andrews, Bonta, and Wormith report five studies indicating that changes in risk level at follow-up assessments were related to expected changes in subsequent recidivism rates.<sup>48</sup> For example, those whose risk scores increased from the first assessment to the second assessment had higher rates of recidivism than those whose scores remained low. However, one of the studies also found that after the first reassessment, additional reassessments added limited improvement to overall predictive validity, suggesting that additional research is needed to fully understand when and how often reassessment is warranted.<sup>49</sup>

**Reliability.** Reliability values for the LS instruments are available for the consistency between raters’ scores, the stability of an

---

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

individual's score across short periods of time, and the consistency with which the items measure the same dimension.

Andrews and Bonta report interrater reliability scores for the LSI-R ranging from  $r=.87$  to  $.94$  when the ratings took place within two months or less.<sup>50</sup> Test-retest reliability ranged from  $r=.95$  to  $.99$  when the instrument was administered by the same rater twice in under a month.<sup>51</sup> For internal consistency, the overall alpha value ranged from  $.64$  to  $.94$  with an average of  $.84$  across 13 studies.<sup>52</sup> The alpha coefficients for each subcomponent varied considerably as measured in the 13 studies, ranging from an average of  $.43$  to  $.78$ .

For the LS/CMI, Andrews, Bonta, and Wormith cite a combined interrater and test-retest reliability of  $r=.88$  for an average interval of 26 days between ratings.<sup>53</sup> For internal consistency, the overall alpha value ranged from  $.86$  to  $.92$  with an average of  $.89$  across eight studies.<sup>54</sup> As with the LSI-R, the alpha coefficients for each subcomponent varied considerably as measured across ten studies, ranging from an average of  $.44$  to  $.80$ .

**Potential for bias: gender.** Some researchers argue that some females follow gender-specific pathways to crime and that the gender-neutral LSI-R, developed on samples of primarily male offenders, has poor predictive validity for those types of females.<sup>55</sup> Reisig and his colleagues, for example, report that the LSI-R predicted recidivism for “economically motivated” female offenders (those similar to male offenders) but not for those who followed gendered pathways to crime in a sample of

women under community supervision in Minnesota and Oregon.<sup>56</sup>

However, the developers of the LSI-R claim that the tool is as reliable and as accurate in the prediction of reoffending for females as with males. They hold that the LS instruments were developed based on a general personality and cognitive social learning perspective of criminality and include separate norms for interpreting male and female total scores.<sup>57</sup> In addition, they cite several evaluations demonstrating the instruments' comparability in predicting male and female recidivism.<sup>58</sup> For example, a published, independent meta-analysis of 25 studies on a total of 14,737 female offenders did not uncover evidence of systematic gender bias in the predictive validity of the LSI-R, showing an average  $r = .35$  for women across these studies.<sup>59</sup> Sixteen of the 25 studies permitted a comparison of the LSI-R's predictive validity by gender; results of this analysis showed that the tool performed comparably for women ( $r_s = .27-.28$ ) and men ( $r_s = .24-.26$ ).

Van Voorhis and her colleagues found that the gender-neutral LSI-R assessment was strongly associated with new arrests in two samples of female probationers in Maui (AUC =  $.72$ ) and Minnesota (AUC =  $.71$ ).<sup>60</sup> However, they also noted that predictive validity increased when the LSI-R was supplemented with gender-responsive factors (AUC =  $.74$  for both sites). The specific factors adding to the improved validity differed somewhat for the two jurisdictions. The authors also found that some of the gender-responsive factors were more related to recidivism than the LSI-R factors, suggesting that treatment priorities

---

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

for females might differ if using the gender-responsive supplement.

Andrews and his colleagues explored the predictive validity of each of the eight LS/CMI factors across five data samples and found each factor predictive of both male and female recidivism.<sup>61</sup> The only significant difference was the enhanced predictive validity of the substance abuse factor for females (AUC = .77 for females and .61 for males). However, they did find that the recidivism rates of low-risk females were substantially lower than those of low-risk men, prompting them to call for an exploration of different cut-off scores that would increase the number of women and decrease the number of men in low risk categories.

The LS test developers note that several gender-informed factors related to education/employment, family/marital (e.g., family conflict), and substance abuse already were in the LSI-R and were carried over to the LS/CMI.<sup>62</sup> In addition, the LS/CMI includes gender-informed items in Section 4: Other Client Issues and Section 5: Special Responsivity Considerations to assist in the development of effective case management plans. In their meta-analysis, Olver and his colleagues found that the LS total scores predicted general recidivism about equally well for men ( $r_w = .30$  and  $.30$ ) and women ( $r_w = .35$  and  $.31$ ). However, they also found that men tended to score higher on areas concerning “antisocial peers, lack of prosocial leisure activities, and substance abuse concerns linked to crime,” whereas women tended to score higher on areas concerning “more serious personal/emotional concerns, financial

problems, and family/marital difficulties” and faced “greater accommodation and education/employment concerns.”<sup>63</sup>

Authors encouraged careful consideration of possible gendered pathways to crime as part of a thorough case planning and program development process.

In sum, there is evidence demonstrating the predictive validity of the LS instruments for female offenders in general. However, variation in LS performance across jurisdictions as well as for specific types of female offenders in addition to the potential utility of individual gender-specific factors as supplements indicate the importance for additional research and local validation of the instruments to ensure their effectiveness.

**Potential for bias: race.** The LSI-R *U. S. Norms Manual Supplement* reports that race/ethnicity had no effect on the total LSI-R scores of community offenders (both male and female) and had a significant, though small, main effect (1% -2% of variability in scores) for inmates.<sup>64</sup> All of the analyses (male, female, community offenders, and inmates) compared Caucasian and African Americans except for male inmates which also included Hispanic, Asian, and Native American offenders. Olver and his colleagues reported significantly smaller effect sizes among ethnic minority offenders ( $r_w = .23$  and  $.23$ ) than among non-minorities ( $r_w = .32$  and  $.29$ ), but concluded that these differences were too small in magnitude to be substantively meaningful.<sup>65</sup>

The findings of additional studies vary. For example, in a study involving 445 African American and Hispanic male inmates released into halfway houses in New Jersey, the predictive validity for rearrest within two



---

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

years was  $r = .08$  for African Americans and  $r = .02$  for Hispanic offenders.<sup>66</sup> The predictive validity for reconviction within two years was  $r = .11$  for African American offenders and  $r = .04$  for Hispanic offenders. The researchers found these correlations low compared to other published studies and concluded that “further analysis of the use of the LSI-R on minority offender populations is warranted and encouraged.”<sup>67</sup>

Another study examined the predictive validity of the LSI-R for a sample of 696 male offenders (72% African American, 15% Hispanic, 13% Caucasian) released from prison in New Jersey.<sup>68</sup> The outcome measure was rearrest within one year. Predictive validity was best for African American offenders (AUC = .61) followed by Caucasian offenders (AUC = .55) and then Hispanic offenders (AUC = .54). The researchers found that African Americans were more likely to be overclassified (rearrest predicted but did not occur), and Hispanics and Caucasians were more likely to be underclassified (no rearrest predicted but rearrest occurred). An additional study reported an overall trend toward more overclassification *and* underclassification for African Americans in a sample of 532 male residents at a federal community corrections center.<sup>69</sup> The sample was 52% African American, 33% Caucasian, and 12% Hispanic. The extent of classification errors varied by the cutoff score and performance measure (i.e., program success or disciplinary incidents) used. The author noted that the low base rate for program failure (11%) and potential reliability issues in scoring the LSI-R may have influenced the results. He concluded that the results highlighted the

need for correctional facilities to validate the instrument on their own populations.

Another study examined the predictive validity of the LSI-R for Native Americans.<sup>70</sup> The study followed 403 community-supervised offenders (56% White and 35% Native American) in the northern midwest for 17 months. The researchers reported predictive validity values of  $r = .18$  for all the offenders,  $r = .23$  for the White offenders, and  $r = .11$  for Native American offenders. Predictive validity was lowest for Native American females with an  $r = -.13$ ; predictive validity for male Native Americans was  $r = .19$ . The researchers suggested additional research to determine whether there are (a) more relevant factors for predicting antisocial behavior among Native Americans, (b) different results when stronger outcome measures (e.g., reconviction rather than rearrest) are used, and (c) different results with a larger sample of Native American women than the current sample of 40. They also questioned whether differences in responsivity factors among Native Americans and the race/ethnicity of the professionals conducting the assessments might affect assessment results.

As with gender, these studies highlight the need for additional research on the predictive validity of the LS instruments for various racial and ethnic groups as well as the importance of validating the instrument for use in specific settings.

***Independent validation.*** The LSI-R has been independently validated across multiple studies and jurisdictions as noted under the “predictive validity” section above. A study investigating the variability in the magnitude of predictive validity estimates

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

found that larger estimates are associated with studies involving LS authors, those conducted in Canada, and those with longer follow-up periods.<sup>71</sup> The authors suggest that the findings are due, in part, to the integrity with which the instruments are used. This explanation was supported in a study by Flores and his colleagues who found that the predictive validity of the LSI-R increased with formal staff training and agency experience with the tool.<sup>72</sup>

### PRACTICAL CONSIDERATIONS.

---

**Vendor and instrument cost.** The LS tools are available for purchase from Multi-Health Systems (MHS).<sup>73</sup>

**Menu of other services.** MHS offers a wide array of services, training, and technical assistance to support the use of LS instruments.

- **IT SERVICES.** Software is available through MHS for completing and scoring the LS instruments. The software can be purchased on a per-use basis, site-licensed, or customized to fit with a jurisdiction's existing database. For larger, jurisdiction-wide implementation, MHS recommends using the Software Developer's Kit (SDK) to integrate the LS tool into the jurisdiction's case management system.<sup>74</sup> This option allows the data to be stored in-house and accessed at any time and requires no maintenance, administration, or technical assistance fees.<sup>75</sup>
- **TECHNICAL ASSISTANCE.**<sup>76</sup> For jurisdictions that opt to use the SDK, MHS has a team of programmers available to help local programmers

incorporate the LS instrument into their case management systems. Toll free assistance is available for those using the standard software package (Smartlink) as well.

Additional assistance is available from a team of researchers to answer questions about the psychometric properties of the instruments. MHS also maintains a Community of Users listserv which serves as a forum to ask questions and share information on policies, procedures and practices.<sup>77</sup> LS users also can submit questions about the tools to the instrument developers.

- **VALIDATION SERVICES.** MHS will norm the LS instruments on the local population once 1,000 assessments have been conducted. There is no additional cost for this service.<sup>78</sup>
- **USER TRAINING.** MHS maintains a training network of certified LS trainers who offer employee training and train-the-trainer programs. The latter saves a jurisdiction the cost of bringing in an outside trainer for each new employee and booster training program. Training costs vary. The jurisdiction submits a request to the network, and trainers bid on the request given the jurisdiction's budget, timing, and needs.<sup>79</sup>

**User qualifications.** To be qualified, test administrators must be trained by an MHS-approved trainer or training program unless they have completed graduate level courses in tests/measurement or can document similar training. Test administrators who do not meet the qualifications must be supervised by a qualified administrator.<sup>80</sup>

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

**Administration time.** The MHS website lists the administration time as 30-45 minutes for the LSI-R and 20-30 minutes for the LS/CMI.<sup>81</sup> The test developers estimate that the client interview can take an hour to an hour and a half.<sup>82</sup>

**Modes of administration.** Information for the LS instruments is collected through a structured interview with the offender, reviews of files and official records, interviews with collaterals such as family members, and, if available, psychological test data.<sup>83</sup>

**Quality assurance.** Several quality assurance considerations follow.

- **OVERRIDE POLICY.** The LS instruments allow professionals to override the quantitative assessment if they identify factors they think deserve special consideration in determining the offender's risk level.<sup>84</sup> If the override option is used, the test administrator is required to provide a written explanation for changing the initial score. The LS/CMI manual notes that overrides are expected in fewer than 10% of cases.<sup>85</sup> The manual also indicates that aggravating and mitigating factors identified in other sections of the LS/CMI may be used to inform and justify the override decision.<sup>86</sup> However, additional research by Wormith and his colleagues indicates that predictive validity decreased when the override was used.<sup>87</sup> The authors found that test administrators used the override much more frequently to increase an offender's risk level than to decrease it and cautioned against the overuse of the practice.

- **FIDELITY.** Because LS instruments require administrator expertise to properly score, formal training is critical to the effective implementation of the instrument. The test developers recommend that initial training be supplemented with periodic booster sessions and audit checks of test administrators' assessments.<sup>88</sup> Agencies with automated databases also can look for systematic trends (e.g., frequent use of overrides for certain types of offenders) in scoring that suggest the need for consultation and/or additional training.<sup>89</sup> Agency staff also uses the MHS Community of Users listserv to discuss quality assurance issues and share strategies to monitor quality.<sup>90</sup>
- **INSTRUMENT REVALIDATION.** The test developers do not have recommendations for the frequency with which LS instruments should be revalidated for a jurisdiction, noting that it depends on an agency's workload and resources. They revalidate the instrument in Ontario approximately every five years and suggest that as a general rule of thumb.<sup>91</sup>

### ENDNOTES

---

<sup>1</sup> See p. 47 in Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct, (4th ed.)*. Cincinnati: Anderson.

<sup>2</sup> See p. 47 in Andrews & Bonta (2006) at endnote 1.

<sup>3</sup> See pp. 22-23 and 151 in Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *The Level of Service/Case Management Inventory (LS/CMI): User's manual*. Toronto: Multi-Health Systems

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

<sup>4</sup> See Andrews & Bonta (2006) at endnote 1, p. 48.

<sup>5</sup> See Andrews & Bonta (2006) at endnote 1, p. 48.

<sup>6</sup> See Andrews & Bonta (2006) at endnote 1, p. 283.

<sup>7</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 155.

<sup>8</sup> See Andrews & Bonta (2006) at endnote 1, p. 48.

<sup>9</sup> See Andrews & Bonta (2006) at endnote 1, p. 48.

<sup>10</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 11.

<sup>11</sup> S. Wormith, personal communication, February 8, 2014.

<sup>12</sup> Andrews, D. A. (1982). *The Level of Supervision Inventory (LSI): The first follow-up*. Toronto: Ontario Ministry of Correctional Services.

<sup>13</sup> Andrews, D. A. & Bonta, J. (1995). *The Level of Service Inventory—Revised: User’s manual*. Toronto: Multi-Health Systems.

<sup>14</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3.

<sup>15</sup> Andrews, D. A., Bonta, J., & Wormith, J. S. (2010). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of Violence risk assessment* (pp. 199-225). New York: Taylor and Francis Group.

<sup>16</sup> See Andrews & Bonta (1995) at endnote 13, p. viii and Andrews, Bonta, & Wormith (2004) at endnote 3, p. 7.

<sup>17</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15.

<sup>18</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 1.

<sup>19</sup> See Andrews & Bonta (1995) at endnote 13, p. 1 and Andrews, Bonta, & Wormith (2004) at endnote 3, p. 1.

<sup>20</sup> J. S. Wormith (personal communication, September 19, 2012).

<sup>21</sup> See Andrews (1982) at endnote 12, p. 2 and Andrews, Bonta, & Wormith (2010) at endnote 15, p. 205.

<sup>22</sup> See Andrews (1982) at endnote 12.

<sup>23</sup> Subsequent factor analyses conducted on the LSI-R produced inconsistent results regarding the instrument’s underlying constructs. Various studies have yielded one, two, and three-factor results. All included a factor related to the propensity to engage in crime. See Andrews, Bonta, & Wormith (2010) at endnote 15, p. 211.

<sup>24</sup> See Andrews & Bonta (1995) at endnote 13, p. 13.

<sup>25</sup> See p. 3 in Andrews, D. A., & Bonta, J. L. (2003). *LSI-R: U. S. norms manual supplement*. North Tonawanda, NY: Multi-Health Systems.

<sup>26</sup> See Andrews & Bonta (2003) at endnote 25, p. 3.

<sup>27</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, p. 205 and Andrews, Bonta, & Wormith (2004) at endnote 3, p. 2.

<sup>28</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 3.

<sup>29</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. xiv. Information also provided by J. Stephen Wormith (personal communication, September 19, 2012).

<sup>30</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. xiv.

<sup>31</sup> The LS/CMI manual (see endnote 3) also reports North American norms. However, the test developers no longer recommend using those norms. The numerous differences observed between the Canadian and U. S. samples make it difficult to interpret a score based on the pooled samples. S. Wormith, personal communication, May 15, 2014.

<sup>32</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, pp. 169, 175. The differences in the n sizes across tables is due to cases in which only total scores or a designated risk level without item data were or were not included. J. S. Wormith, personal communication, May 15, 2004.

<sup>33</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p.3.

<sup>34</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 3.

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

<sup>35</sup> See Andrews & Bonta (2003) at endnote 25, p. 4. The cutoff scores differ for institutional classification.

<sup>36</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 36.

<sup>37</sup> See Andrews & Bonta (1995) at endnote 13, p. 14.

<sup>38</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 36.

<sup>39</sup> The authors also reported significant differences in effect size magnitude by geographic region, which was highest in Canadian samples ( $r_w = .38$  and  $.43$  for fixed- and random-effects models, respectively), followed by non-North American samples ( $r_w = .30$  and  $.29$ ), and United States samples ( $r_w = .20$  and  $.22$ ). Olver, M., Stockdale, K., & Wormith, J. (2014). Thirty years of research on the Level of Service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment, 26*, 156-176.

<sup>40</sup> Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation, 72* (3), 22-29.

<sup>41</sup> See pp. 24-25 in Vose, Cullen, & Smith (2008) at endnote 40. The range reported excludes studies examining program completion and parole violations.

<sup>42</sup> Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*, 575-607. The effect size adjusted for sample size is  $r = .33$ .

<sup>43</sup> Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the “unparalleled” measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior, 29*, 397-426. The effect size adjusted for sample size is  $.39$ . The study also provided a mean effect size of  $r = .26$  ( $r = .28$ , adjusted for sample size) for predicting violent recidivism. A subsequent meta-analysis by Campbell and her colleagues reports a mean effect size for violent recidivism of  $r = .25$  ( $r = .28$ , adjusted for sample size). See Campbell, M. A., French, S.,

& Gendreau (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36*, 567-590.

<sup>44</sup> Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency, 52*, 7-27. The authors also report a mean effect size for violent recidivism of  $r = .25$ .

<sup>45</sup> For the federal probation study, see Flores, A. W., Lowenkamp, C. T., Smith, P., & Latessa, E. J. (2006). Validating the Level of Service Inventory—Revised on a sample of federal probationers. *Federal Probation, 70*(2), 44-49. For the Iowa probation and parole study, see Lowenkamp, C. T., & Bechtel, K. (2007). The predictive validity of the LSI-R on a sample of offenders drawn from the records of the Iowa Department of Corrections data management system. *Federal Probation, 71*(3), 25-29.

<sup>46</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, p. 213. This profile of the LS/CMI is focused on Section 1 “General Risk/Need Factors.” The LS/CMI manual (see pp. 122-123 in Andrews, Bonta, & Wormith, 2004, at endnote 3) also provides limited information about the predictive validity of Sections 2 through 5.

<sup>47</sup> Girard, L., & Wormith, J. S. (2004). The predictive validity of the Level of Service Inventory—Ontario Revision on general and violent recidivism among various offender groups. *Criminal Justice and Behavior, 31*, 150-181. The effect size for violent conviction is  $r = .28$ .

<sup>48</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, pp. 213-214.

<sup>49</sup> Arnold, T. (2007). *Dynamic changes in the Level of Service Inventory-Revised (LSI-R) scores and the effects on prediction accuracy*. Master’s thesis, St. Cloud University, St. Cloud, MN.

<sup>50</sup> See Andrews & Bonta (1995) at endnote 13, p. 35.

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

- <sup>51</sup> See Andrews & Bonta (1995) at endnote 13, p. 35.
- <sup>52</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, pp. 206-208.
- <sup>53</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, pp. 114-115.
- <sup>54</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, pp. 206-208.
- <sup>55</sup> See, for example, Holtfreter, K. & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women. *Journal of Contemporary Criminal Justice*, 23, 363-382; Reisig, M. D., Holtfreter, K., & Morash, M. (2006). Assessing recidivism risk across female pathways to crime. *Justice Quarterly*, 23, 384-405.
- <sup>56</sup> See Reisig, et al. (2006) at endnote 55, p. 397. The predictive validity was  $r = .24$  for economically motivated offenders and  $r = -.13$  for gendered pathway offenders.
- <sup>57</sup> See Andrews & Bonta (1995) at endnote 13, p. 48 and Andrews, Bonta, & Wormith (2004) at endnote 3, p. 143.
- <sup>58</sup> Andrews, D. A., Bonta, J., & Wormith, J. S. (2009). *LS/CMI: A gender-informed risk/need/responsivity assessment*. North Tonawanda, NY: Multi-Health Systems.
- <sup>59</sup> Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy*, 8, 183-208.
- <sup>60</sup> Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior*, 37, 261-288.
- <sup>61</sup> Andrews, D. A., Guzzo, L., Raynor, P., Rowe, R. C., Rettinger, L. J., Brews, A., & Wormith, J. S. (2012). Are the major risk/need factors predictive of both female and male reoffending? A test with the eight domains of the Level of Service/Case Management Inventory. *International Journal of Offender Therapy and Comparative Criminology*, 56(1), 113-133.
- <sup>62</sup> Andrews, D. A., Bonta, J., & Wormith, J. S. (2009). *Level of Service/Case Management Inventory (LS/CMI). Supplement: A gender-informed risk/need/responsivity assessment*. North Tonawanda, NY: Multi-Health Systems.
- <sup>63</sup> Olver et al. (2013) at endnote 39, p. 14.
- <sup>64</sup> See Andrews & Bonta (2003) at endnote 25, p. 10.
- <sup>65</sup> Olver et al. (2013) at endnote 39.
- <sup>66</sup> Schlager, M. D., & Simourd, D. J. (2007). Validity of the Level of Service Inventory—Revised (LSI-R) among African American and Hispanic male offenders. *Criminal Justice and Behavior*, 34, 545-554.
- <sup>67</sup> See Schlager & Simourd (2007) at endnote 66, p. 553.
- <sup>68</sup> Fass, T. L., Heilbrun, K., Dematteo, D., & Fretz, R. (2008). The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice and Behavior*, 35, 1095-1108.
- <sup>69</sup> Whiteacre, K.W. (2006). Testing the Level of Service Inventory-Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review*, 17, 330-342.
- <sup>70</sup> Holsinger, A. M., Lowenkamp, C. T., & Latessa, E. J. (2006). Exploring the validity of the Level of Service Inventory-Revised with Native American offenders. *Journal of Criminal Justice*, 34, 331-337.
- <sup>71</sup> Andrews, D. A., Bonta, J., Wormith, J. S., Guzzo, L., Brews, A., Rettinger, J., & Rowe, R. (2011). Sources of variability in estimates of predictive validity: A specification with Level of Service general risk/need. *Criminal Justice and Behavior*, 38, 413-432.
- <sup>72</sup> Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity. *Journal of Criminal Justice*, 34, 523-529. Also see Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, DC: The Institute on Crime,

## Appendix: RNA Instrument Profile for the LSI-R and LS/CMI

---

Justice and Corrections at the George Washington University. Austin and his colleagues found that reliability increased after additional training was provided to the test administrators.

<sup>73</sup> Information on the LSI-R is available at <http://www.mhs.com/product.aspx?gr=saf&prod=lsi-r&id=overview>; and information on the LS/CMI is available at <http://www.mhs.com/product.aspx?gr=saf&prod=ls-cmi&id=overview>. According to an undated report provided by Tammy Howell of MHS, substantial discount rates are available for states and counties interested in adopting LS instruments. See, Howell, T. (n.d.). *LSI-R, LS/RNR, and LS/CMI documentation*. North Tonawanda, NY: Multi-Health Systems. Retrieved from <http://www.scstatehouse.gov/archives/citizeninterestpage/SentencingReformCommission/SentencingReform.php>

<sup>74</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, p. p. 200.

<sup>75</sup> See Howell (n.d.) at endnote 73: MHS charges a one-time fee to purchase the SDK and charges an annual fee based on the number of assessments conducted. The per assessment fee is based on a sliding scale according to the number of assessments purchased. A standard software package (Smartlink) also is available.

<sup>76</sup> Information for this section comes from Howell (n.d.) at endnote 73.

<sup>77</sup> The Community of Users also was discussed by the test developers: J. Bonta and S. Wormith, personal communication, April 17, 2012.

<sup>78</sup> J. Bonta and S. Wormith, personal communication, April 17, 2012. Also see Howell (n.d.) at endnote 73.

<sup>79</sup> See Howell (n.d.) at endnote 73 and the MHS web site at [https://ecom.mhs.com/\(S\(go2ocy45brgsoq55mnrk5m55\)\)/saf\\_om.aspx?id=Training](https://ecom.mhs.com/(S(go2ocy45brgsoq55mnrk5m55))/saf_om.aspx?id=Training).

<sup>80</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 6.

<sup>81</sup> See MHS website at [https://ecom.mhs.com/\(S\(oazaozbybrdg5uz5fc1yzj55\)\)/product.aspx?gr=saf&prod=lsi-r&id=overview](https://ecom.mhs.com/(S(oazaozbybrdg5uz5fc1yzj55))/product.aspx?gr=saf&prod=lsi-r&id=overview) and [https://ecom.mhs.com/\(S\(2s1oyojq5lrrttayrndkko45\)\)/product.aspx?gr=saf&prod=ls-cmi&id=overview](https://ecom.mhs.com/(S(2s1oyojq5lrrttayrndkko45))/product.aspx?gr=saf&prod=ls-cmi&id=overview).

<sup>82</sup> J. Bonta and S. Wormith, personal communication, April 17, 2012.

<sup>83</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, p.200.

<sup>84</sup> See Andrews & Bonta (1995) at endnote 13, p. 12 and Andrews, Bonta, & Wormith (2004) at endnote 3, p. 4.

<sup>85</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 4.

<sup>86</sup> See Andrews, Bonta, & Wormith (2004) at endnote 3, p. 123.

<sup>87</sup> Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior*, 39, 1511-1538. The authors found that predictive validity decreased for all offenders but was particularly lowered for sexual offenders.

<sup>88</sup> See Andrews, Bonta, & Wormith (2010) at endnote 15, pp. 209-210.

<sup>89</sup> J. Bonta and S. Wormith, personal communication, April 17, 2012.

<sup>90</sup> J. Bonta and S. Wormith, personal communication, April 17, 2012.

<sup>91</sup> J. Bonta and S. Wormith, personal communication, April 17, 2012.

---

# The Offender Screening Tool (OST)

---

## OST GLOSSARY OF TERMS

---

<b>Risk</b>	Authors adhere to Andrews and Bonta’s risk principle, stating that "supervision strategies should prioritize treatment and probation resources for higher risk offenders." <sup>1</sup>
<b>Static risk</b>	Authors use this term to describe risk factors that "contribute to an individual's risk to reoffend but cannot be changed." <sup>2</sup>
<b>Dynamic risk</b>	In combination with static (historical) risk factors, dynamic (changeable) risk factors have been found to be significant predictors of recidivism. The authors state that dynamic factors help "identify potential targets for treatment" and "contribute to an individual's overall risk to reoffend." <sup>3</sup>
<b>Needs</b>	Authors adhere to Andrews and Bonta’s needs principle, stating that "probation strategies should target interventions to criminogenic needs. Supervision should address the offenders’ needs that are directly linked to criminal behavior." <sup>4</sup>
<b>Responsivity</b>	Authors acknowledge Andrews and Bonta’s responsivity principle, stating that "probation staff should be responsive to temperament, learning style, motivation, culture, and gender when assigning programs". <sup>5</sup>
<b>Protective factors</b>	Term not used.
<b>Strengths</b>	Term not used.
<b>Recidivism</b>	In the independent assessment of OST <sup>6</sup> , the evaluators used five separate measures of recidivism: (1) petition to revoke, (2) petition to revoke with new arrest, (3) revoked, (4) any arrest, and (5) felony arrest.

---

## HISTORY & CURRENT USE.

---

**Creation.** In 1996, the Maricopa County Adult Probation Department (MCAPD) in Arizona reviewed existing offender assessment practices as part of its commitment to research-based practices.<sup>7</sup> This review prompted the MCAPD to seek a more meaningful offender risk and needs assessment tool. In response, the staff of MCAPD, with the assistance of research

consultant Dr. David Simourd, developed the Offender Screening Tool (OST) in 1998.

**Current use.** Although developed for use in Maricopa County, the OST was subsequently validated on the probation population statewide and was adopted by the Arizona Administrative Office of the Courts (AOC) for statewide use with probationers in January, 2005.<sup>8</sup> Prior to the OST, most counties in the state were using variants of



## Appendix: RNA Instrument Profile for the OST

---

the Wisconsin risk and needs assessment tool, which had never been validated for the Arizona probation population.<sup>9</sup> In addition, there was evidence that probation officers across the state were not using the instruments consistently, nor were the results being used to inform decisions about the level of services to be received.<sup>10</sup>

The OST is also used in local probation departments in Virginia with misdemeanor offenders.<sup>11</sup>

### DEVELOPMENT.

---

**Instrument purpose.** MCAPD sought an instrument that would assess both risk and needs of the offender using static and dynamic measures directly related to the key predictors of criminal behavior. The goal was to implement an instrument that would gauge the likelihood of individual reoffending and also identify specific offender needs that could be used to inform more effective treatment and service delivery. Additionally, to increase the likelihood that the tool would be used consistently and as intended, MCAPD wanted a tool that probation staff viewed as meaningful.<sup>12</sup>

Rather than draw on a pre-existing risk and needs assessment instrument, MCAPD decided to create its own tool. This decision reflected several factors, including a concern about the annual cost of using an existing proprietary tool given the large number of assessments done each year, the need to identify a tool that was valid for use with the local population of offenders, and a strong desire to involve probation staff in the development of the tool. At that time,

MCAPD had also decided to reengineer the operation of their presentence division. Through reengineering, the presentence process was streamlined, duplicated effort was eliminated, and the OST system was introduced to the department.

With the OST system, probation officers make use of three main assessment tools: a full assessment, a reassessment, and a brief screener. The OST, the full assessment tool, is administered at the presentence phase to identify offender behaviors over the previous 12 to 36 months. Results are used to guide case management decisions. To capture the effect of probationary intervention and inform case management decisions over time, Arizona employs the Field Reassessment Offender Screening Tool (FROST), nearly identical to the OST in items and scoring, to reassess offenders for changes in risk and needs over time.<sup>13</sup> The FROST is designed to be conducted at 6 month intervals. Completing either the full assessment or the reassessment requires a review of the case file and an interview with the offender. Some judgment is needed to score items on the instrument.

An abbreviated version of the OST, the Modified Offender Screening Tool (MOST), was developed for expedited use and draws on 8 items from the OST.<sup>14</sup> Designed as a relatively quick screening tool, higher scores on the MOST are a signal to probation officers to administer the full OST.

**Approach to instrument development.** In creating the OST, the developers used an approach that was more theoretically than statistically driven.<sup>15</sup> From this framework, they incorporated factors related to both risk

## Appendix: RNA Instrument Profile for the OST

---

and needs in a single instrument that they believed reflected the latest thinking about the psychology of criminal conduct. Related in design to the Level of Service Inventory-Revised (LSI-R), OST employs a similar class of variables drawn from correctional and developmental literature and from existing meta-analytic research identifying the strongest predictors of recidivism.<sup>16</sup>

Assessment questions were based on these variables identified in the research literature as related to criminal behavior.<sup>17</sup> Although the developers assert that all factors on the OST are related to recidivism, those categories that are stronger predictors are given more weight (i.e., more items) and therefore have greater influence on the overall risk score.<sup>18</sup>

In addition to this theoretical relationship, instrument developers sought items that showed a statistically significant relationship with recidivism, had face validity to facilitate buy-in from the court community, and that could be easily scored by probation officers to ensure consistent and proper use of the tool. They also sought to include items that were relevant in the treatment process and strongly preferred dynamic over static items. The final OST is comprised of items that are 61% dynamic.<sup>19</sup>

### CONTENT.

---

**Structure.** The OST and FROST each produce a single overall score from a set of nine subscales. This overall score is used to determine the offender's recidivism risk level. Overall scores are positively related to multiple measures of the offender's risk of recidivism, with the two primary outcome

measures being (1) petition to revoke and (2) any new arrest.<sup>20</sup>

When using the full instrument, probation officers are told that the primary needs areas (those that require intervention through case planning) are those identified by the nine subscales. Scores for each of the nine subscales are used to identify and prioritize the offender's needs for case planning and service provision. A 10th section of the OST and FROST contains two additional items, referred to as responsivity factors. They are not criminogenic and are not incorporated into the computation of overall risk or individual needs.<sup>21</sup>

**Items and domains.** The OST is comprised of 42 items across the nine different risk and need subscales (or domains). Each domain is comprised of 2-9 items that may be static or dynamic. The nine domains include: vocational/financial (5 items), education (3 items), family & social relationships (8 items), residence & neighborhood (2 items), alcohol (3 items), drug abuse (3 items), mental health (2 items), attitude (7 items), and criminal behavior (9 items).

The tenth section on responsivity factors includes two additional physical/medical health items (for a total of 44 items) and is used to identify whether or not health-related concerns may pose potential barriers to successful offender treatment.

In general, OST items are scored on the basis of patterns of behavior rather than a single incident (e.g., a single incident of alcohol use should not necessarily be coded as problematic use). Probation officers are encouraged to have at least one or two

## Appendix: RNA Instrument Profile for the OST

---

reasons that explain why each OST item is scored as it is for the offender.

**Reporting and cutoffs.** When first introduced, the OST categorized offenders into one of three levels of risk (low, moderate, high) based on overall scores. The cutoff values used to create these three categories were estimated based on the cutoff scores used in the LSI-R. Following an initial period of use, these cutoffs were revised based on actual OST data from the local probation population in Maricopa County.<sup>22</sup>

When the OST was adopted statewide, the cutoff scores were reexamined. Results from a 2008 independent statewide validation study indicated that the range of scores in the moderate risk category was too large to sufficiently differentiate offenders.<sup>23</sup> Based on these findings, the OST risk categories were again revised, this time expanding from three risk categories to four. In addition, separate cutoff values were established for men and women. These new cutoff values were as follows: For males, low (1-5 points), moderate (6-10 points), moderate-high (11-17 points), and high (18+ points); for females, low (0-8 points), moderate (9-13 points), moderate-high (14-20 points), and high (21+ points). The low-risk cutoff values were selected to align with a 15 percent failure (recidivism) rate.

Unlike some tools, the OST does not produce similar ranking categories to identify level of need in each domain. Rather, probation officers are encouraged to target needs identified by dynamic items in high-scoring domains.

### INSTRUMENT RELIABILITY AND VALIDITY.

---

**Populations studied.** Following the creation of OST on a construction sample of male and female Maricopa County probationers,<sup>24</sup> the instrument was validated in 2003 on a statewide sample of male and female probationers<sup>25</sup> and independently validated on another Arizona statewide sample in 2008.<sup>26</sup> A statewide validation study has also been completed in Virginia.<sup>27</sup> In these validation studies, researchers selected representative samples of offenders who had a case closed within a suitable timeframe to allow for an evaluation of probation outcome (e.g., at least six months). The validation study in Virginia, for example, distinguished offenders based on age, sex, race/ethnicity, criminal history, current charge, and geographic location.

**Predictive validity.** The developers found that prior OST risk scores were significantly higher for offenders whose current probation status was deemed "unsatisfactory" vs. those whose current behavior was found to be "satisfactory."<sup>28</sup>

A more rigorous analysis of the OST's predictive validity was undertaken in 2008.<sup>29</sup> With respect to recidivism, researchers found that the OST works best in Arizona as a predictor of petitions to revoke ( $r=.23$ ) and less well as a predictor of any arrest ( $r=.12$ ).<sup>30</sup> In Virginia, one outcome variable was examined: probation closure type.<sup>31</sup> Closure type was coded as a) successful, b) transfer in-sent back, or c) unsuccessful. A linear relationship between OST scores and outcome was expected because greater OST scores are designed to be reflective of greater

## Appendix: RNA Instrument Profile for the OST

---

criminal need issues. A statistically significant relationship between the OST score and outcome was found ( $r=.19$ ).

**Reliability.** Lowenkamp and colleagues reported levels of inter-rater agreement in Arizona above 90% for 25 of the 42 OST items.<sup>32</sup> Lower percentages of agreement tended to emerge from items that required the assessor to count or identify times of occurrences (e.g., two or less times unemployed,) and for items that required more professional discretion (e.g., client being in denial about alcohol use). The evaluators recommended that rater consistency could be improved with more training.

**Potential for bias.** Simourd examined the differential validity of the OST on males and females and found no significant differences in overall scores, but significant differences within certain domains of the tool.<sup>33</sup> Males were found to have significantly greater scores on the Education, Alcohol, and Criminal Behavior domains, while females had significantly greater scores on Vocational/Financial, Family and Social Relationships, and Mental Health domains. He found no significant differences by county or by type of offense. Simourd concluded that the observed gender differences were small in practical terms and therefore made no recommendations for change.

Lowenkamp and colleagues examined the differential validity of OST on sex and ethnicity (Hispanic vs. non-Hispanic) and concluded that the tool performs adequately for all subgroups in predicting petitions to revoke, but less well for other measures of

recidivism.<sup>34</sup> They suggested altering cutoff scores to improve predictive validity but did not make any further recommendations. Following the evaluation, the OST moved to four categories of risk and established different cutoff values for men and women.

**Independent validation.** One independent validation has been conducted to date.<sup>35</sup>

### PRACTICAL CONSIDERATIONS.

---

**Vendor and instrument cost.** The OST system is non-proprietary. For more information, contact Dr. Jennifer Ferguson of MCPAD ([jferguso@apd.maricopa.gov](mailto:jferguso@apd.maricopa.gov)) or Dr. David Simourd of ACES Inc. ([dave@acesink.com](mailto:dave@acesink.com)).

**Menu of other services.** Not applicable for this non-proprietary tool, although independent consultants have offered research support and validation services.

**User qualifications.** The OST is administered by the Arizona Adult Probation Department (APD) presentence division. Individual probation officers administer the reassessment (FROST).<sup>36</sup>

Following the 2008 reliability and validation study,<sup>37</sup> APD instituted mandatory initial and refresher training requirements for presentence division staff and probation officers. All probation officers are trained on the instrument. Presentence screeners receive training on interviewing skills and, after completing several interviews in the field, participate in focus groups to exchange feedback and refine their OST administration skills. Probation officers must complete a three-year refresher training, which includes a review the OST system and

## Appendix: RNA Instrument Profile for the OST

---

addresses the topic of developing appropriate case plans.<sup>38</sup> In addition, the training program reviews the OST and FROST Scoring Guides, which provide descriptions of and scoring tips for all items.

**Administration time.** Developers say the OST and FROST take, on average, about 25 minutes to complete.<sup>39</sup>

**Modes of administration.** Information used to complete the OST and FROST is drawn from a structured interview that relies partly on offender self-report. The administering presentence screener or probation officer leads the interview. The computerized OST system automatically calculates assessment results.

**Quality assurance.** When adopting any offender assessment tool, jurisdictions must be prepared to ensure appropriate use and maintenance over time. Protocols established by Maricopa County and the state of Arizona Probation Departments are described below.

- **OVERRIDE POLICY.** The stated goal in Arizona is to minimize the number of overrides to the OST recommendations.<sup>40</sup> When first implemented, the developers indicated that an override of OST results should occur in no more than 10% of cases. Currently, there is no specific numerical target and no systematic effort to track overrides. The decision to override the instrument recommendation is made on a case-by-case basis when the probation officer believes it is justified.
- **FIDELITY.** Reliability in the use of the instrument depends to a great extent on

training. In Maricopa County, a refresher training system has been developed to improve scoring consistency among presentence screeners and probation officers. These users first view an educational refresher training video online and then complete a scoring test. If the user does not pass the scoring test, they are required to attend an in-person classroom refresher training course and retake the scoring test. If the user still does not meet internal quality control standards after completing the classroom course, their supervisor incorporates training into their performance evaluation plan.<sup>41</sup>

APD has instituted other mechanisms to ensure fidelity. In addition to mandatory initial and refresher training programs for presentence division staff and field probation officers (see *User Qualifications*), state presentence screeners are trained to perform quality control checks on the information gathered from the structured interview with the offender and entered into the automated system (such as by verifying criminal history information provided by the offender with existing records). Moreover, the computerized OST system automates the scoring process and contains built-in mechanisms to ensure that required questions are not skipped to minimize user error.

- **INSTRUMENT REVALIDATION.** In their 2008 independent evaluation, Lowenkamp and colleagues recommend that tests of the instruments' predictive

## Appendix: RNA Instrument Profile for the OST

---

validity should be conducted at least once every three years.<sup>42</sup>

### ENDNOTES

---

<sup>1</sup> See p. 4 in Arizona Adult Probation Services Division (2010, July). *Eight evidence-based principles for effective interventions & timeline of significant events in Arizona probation*. Phoenix, AZ: Author. Retrieved from

[http://www.azcourts.gov/LinkClick.aspx?fileticket=1CL48eRs\\_Nw%03d&tabid=2654](http://www.azcourts.gov/LinkClick.aspx?fileticket=1CL48eRs_Nw%03d&tabid=2654). Risk

principle discussed in Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct*, (4th ed.). Cincinnati: Anderson.

<sup>2</sup> See p. 474 in Ferguson, J. (2002). Putting the “what works” research into practice: An organizational perspective. *Criminal Justice and Behavior*, 29, 472-492. Retrieved from <http://cjb.sagepub.com/content/29/4/472>

<sup>3</sup> See pp. 474-475 in Ferguson (2002) at endnote 2.

<sup>4</sup> See endnote 1.

<sup>5</sup> See endnote 1.

<sup>6</sup> Lowenkamp, C. T., Latessa, E., & Bechtel, K. (2008). *A reliability and validation study of the Offender Screening Tool (OST) and Field Reassessment Offender Screening Tool (FROST) for Arizona*. Cincinnati: Center for Criminal Justice Research, University of Cincinnati.

<sup>7</sup> See Ferguson (2002) at endnote 2.

<sup>8</sup> See Lowenkamp et al. (2008) at endnote 6. See also Simourd, D. (2003). *Arizona Supreme Court: Administrative Office of the Courts, Adult Probation Services Division, Risk and needs assessment project*. Kingston, ON: Algonquin Correctional Evaluation Services.

<sup>9</sup> Baird, S.C. (1981). Probation and parole classification: The Wisconsin Model. *Corrections Today*, 43, 36-41.

<sup>10</sup> See Ferguson (2002) at endnote 2.

<sup>11</sup> Simourd, D. J., (2010, August). *Validation of the Offender Screening Tool (OST) for the Virginia local probation agencies*. Kingston,

---

ON: Algonquin Correctional Evaluation Services.

<sup>12</sup> See Ferguson (2002) at endnote 2.

<sup>13</sup> Arizona Adult Probation Services Division. (2009, July update). *FROST scoring guide*. Phoenix, AZ: Author.

<sup>14</sup> Arizona Adult Probation Services Division. (2009, July update). *MOST scoring guide*. Phoenix, AZ: Author.

<sup>15</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

<sup>16</sup> Andrews, D. A. & Bonta, J. (1995). *The Level of Service Inventory—Revised: User’s manual*. Toronto: Multi-Health System.

<sup>17</sup> See Ferguson (2002) at endnote 2.

<sup>18</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012..

<sup>19</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

<sup>20</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>21</sup> Arizona Adult Probation Services Division. (2009, July update). *OST scoring guide*. Phoenix, AZ: Author.

<sup>22</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

<sup>23</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>24</sup> See Ferguson (2002) at endnote 2.

<sup>25</sup> See Simourd (2003, July) at endnote 8.

<sup>26</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>27</sup> See Simourd (2010, August) at endnote 11.

<sup>28</sup> See Simourd (2003, July) at endnote 8.

<sup>29</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>30</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>31</sup> See Simourd (2010, August) at endnote 11.

<sup>32</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>33</sup> See Simourd (2003, July) at endnote 20.

<sup>34</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>35</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>36</sup> See Arizona Adult Probation Services Division (2009, July update) at endnote 13.

<sup>37</sup> See Lowenkamp et al. (2008) at endnote 6.

<sup>38</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

<sup>39</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

<sup>40</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

## Appendix: RNA Instrument Profile for the OST

---

---

<sup>41</sup> J. Ferguson and D. Simourd, personal communication, March 5, 2012.

<sup>42</sup> See Lowenkamp et al. (2008) at endnote 6.

# Ohio Risk Assessment System (ORAS): Community Supervision Tool (CST)

---

## ORAS GLOSSARY OF TERMS

---

<b>Risk</b>	Authors do not define “risk,” but explain the logic of Andrews and Bonta’s (1994) risk principle: The intensity of programmatic treatment should match the offender’s risk level so that “the most intensive programming should be allocated to moderate- and high-risk cases, while low-risk cases should be allocated little if any programming.” <sup>1</sup>
<b>Static risk</b>	Term not used in either report on the creation of the ORAS. However, authors use this term to describe risk factors that, because of the nature of the item(s), cannot be reduced over time. <sup>2</sup> Also referred to as “past criminal behavior.” <sup>3</sup>
<b>Dynamic risk</b>	Criminogenic or “crime-producing” needs, or “factors that, when changed, have been shown to result in a reduction in recidivism.” <sup>4</sup>
<b>Needs</b>	Authors do not use this term except to discuss criminogenic needs (see dynamic risk, above), but describe the needs principle as suggesting that “effective classification systems should identify dynamic risk factors directly related to recidivism so that they can be used to target programmatic needs.” <sup>5</sup>
<b>Responsivity</b>	Offender issues that “are not directly related to recidivism, but instead have the potential to restrict the efficacy of treatment. [They] are not used in the final calculation of risk, but instead are used as case planning factors that should be addressed to improve likelihood that programming will reduce recidivism.” <sup>6</sup>
<b>Protective factors</b>	Term not used in either report on the creation of the ORAS.
<b>Strengths</b>	Term not used in either report on the creation of the ORAS.
<b>Recidivism</b>	The ORAS Community Supervision Tool (CST) predicts the likelihood that community-based adult offenders will be arrested for a new crime, as measured in a 12-month follow-up period. <sup>7</sup>

---

## HISTORY & CURRENT USE.

---

**Creation.** In 2006, the Ohio Department of Rehabilitation & Correction (ODRC) hired researchers from the University of Cincinnati

(UC) Center for Criminal Justice Research to develop an integrated, automated assessment system of offender risk, needs, and barriers to treatment that could be used



## Appendix: RNA Instrument Profile for the ORAS-CST

---

to better inform decision-making statewide and ultimately reduce recidivism.<sup>8</sup>

**Current use.** The ODRC officially implemented the full ORAS statewide in Ohio as of March 2011, following the completion of construction, validation, and pilot testing studies on the system.<sup>9</sup> Although only recently adopted, a number of other states are using the ORAS, including Connecticut, Colorado, Montana, Nevada, New Hampshire, and Vermont, as well as a number of counties in Florida, Pennsylvania, and California.<sup>10</sup> A version of the complete ORAS was recently validated for statewide use in Indiana as the IRAS (i.e., Indiana Risk Assessment System) and in Texas as the TRAS (i.e., Texas Risk Assessment System).<sup>11</sup> Other studies are also currently planned or underway in Connecticut and Ventura County, California.<sup>12</sup>

### DEVELOPMENT.

---

**Instrument purpose.** The goal in creating the ORAS was to develop a unique, standardized system of offender assessment tools that could be used at various decision points in the criminal justice system to reduce recidivism, and that would facilitate communication and continuity in case management across criminal justice agencies.<sup>13</sup> The ORAS contains four full assessment tools (each designed for use at pretrial, at prison intake, with community supervision populations, or with reentry populations) and two brief screener tools (for use with prison and community supervision populations).<sup>14</sup> The authors have also recently developed a tool specifically for misdemeanants.<sup>15</sup> This profile focuses on the component of the ORAS developed

specifically for use with community-based populations of offenders (i.e., probation, parole, offenders in residential facilities or other community alternatives): the Community Supervision Tool, or CST. ORAS developers recommend administering the full CST, and not the short screening version, if using the results of the tool at the sentencing stage.<sup>16</sup>

**Approach to instrument development.** To create the CST, UC researchers adopted a prospective design.<sup>17</sup> This means that researchers identified current offenders (all adults charged with a criminal offense and referred to probation services during the period of data collection) for participation in the study, interviewed them to collect data on potential risk factors thought to predict recidivism, and observed these offenders over time (one year) to gather recidivism data. Researchers opted for a prospective study rather than a retrospective study which uses historical or archival data from past offenders to create the assessment tools because many potential offender risk factors considered for use in the CST or in other ORAS tools may not have been previously documented by criminal justice agencies. This approach allowed UC researchers to examine a comprehensive battery of over 200 potential risk factors for possible inclusion in the instrument(s).<sup>18</sup>

From this large pool of items, UC researchers eliminated those which failed to show a statistically significant relationship with recidivism. Researchers then conducted factor analyses and scale reliability tests to organize the content of the CST into seven domains or categories and to pare down the tool to the fewest items possible for optimal

## Appendix: RNA Instrument Profile for the ORAS-CST

---

predictive validity. In the item selection process, if a dynamic risk item performed as well or better than a comparable static risk item, UC researchers made a decision to prioritize the inclusion of the dynamic item because of the ability of dynamic items to measure and reflect changes over time. UC researchers indicated that generally, dynamic items were just as predictive as, if not better than, static risk items.<sup>19</sup>

### CONTENT.

---

**Structure.** The CST generates a single overall score from a set of seven subscales. This overall score represents the offender's risk of recidivism. Scores for each of the seven separate subscales of the CST are used to identify and prioritize the offender's needs for case planning and service provision (see *Items and domains* section below for a list of the needs domains addressed by the CST).

A separate section lists responsivity factors as other potential areas of concern that may inform case planning decisions. These factors are not criminogenic and are not incorporated into the computation of risk.<sup>20</sup>

**Items and domains.** The ORAS CST consists of 35 items in 7 subscales: criminal history (6 items); education, employment, and finances (6 items); family and social support (5 items); neighborhood problems (2 items); substance abuse (5 items); antisocial associations (4 items); and antisocial attitudes and behavioral problems (7 items).

The CST also documents the following treatment barriers to inform case planning: low intelligence, physical handicap, reading and writing limitations, mental health issues,

offender motivation to change/participate in treatment, transportation, child care, language, ethnicity, cultural barriers, history of abuse/neglect, and interpersonal anxiety.<sup>21</sup>

**Reporting and cutoffs.** The ORAS CST groups offenders into four levels of risk (low, moderate, high, very high) based on their overall score. The cutoff scores differ by gender.

The ORAS CST also groups offenders, on each subscore, into three priority levels (low, moderate, high) to inform decisions about which offender needs should be prioritized in case planning and service provision. Offenders categorized as "high" in a particular domain are more likely to reoffend. The cut points vary by domain, but not by gender.

All cutoff scores are identified in the ORAS manual.<sup>22</sup>

### INSTRUMENT RELIABILITY AND VALIDITY.

---

**Populations studied.** In addition to the statewide Ohio creation and validation samples of probation-eligible male and female adult offenders, Indiana has also completed a statewide validation study of the tool (report forthcoming).<sup>23</sup>

**Predictive validity.** ORAS developers reported a correlation of  $r = .36$  between ORAS CST risk level and recidivism in the Ohio study. Moreover, case management priority levels for each of the 7 subscale domains also correlated individually with recidivism (criminal history,  $r = .20$ ; education and finances,  $r = .22$ ; social support,  $r = .12$ ; neighborhood problems,  $r =$

## Appendix: RNA Instrument Profile for the ORAS-CST

---

.20; substance abuse,  $r = .14$ ; antisocial associates,  $r = .32$ ; and antisocial attitudes,  $r = .24$ ), providing further evidence that these domains identify criminogenic needs.<sup>24</sup>

**Reliability.** No data available at the time of this report but see “Independent validation” section below.

**Potential for bias.** There is little evidence currently available on the issue of bias with the ORAS CST.

- **GENDER.** ORAS developers reported correlations between the ORAS CST risk level and recidivism in the Ohio study of  $r = .37$  for males and  $r = .30$  for females.<sup>25</sup> In general, female offenders tend to produce lower scores on the ORAS than males. Instrument developers established different risk level cutoff scores by gender to reflect this.<sup>26</sup>
- **RACE.** No data currently available.

**Independent validation.** As of this publication, no independent validation studies of the ORAS have been published. However, Texas reportedly has recently completed the first independent interrater reliability study and an independent predictive validity study using a random statewide sample.<sup>27</sup>

### PRACTICAL CONSIDERATIONS.

---

**Vendor and instrument cost.** The ORAS tools are non-proprietary. For more information, contact Ms. Jennifer Luxat UC ([luxjl@ucmail.uc.edu](mailto:luxjl@ucmail.uc.edu)).

**Menu of other services.** UC offers a wide array of services, training, and technical assistance to support ORAS implementation.

- **IT SERVICES.** Customized software is available for purchase. Depending on the level of customization and other options selected, the price of an automated module system currently ranges from \$15K – 100K.<sup>28</sup> As of this report, customization options include:<sup>29</sup>
  - A base module system that is hosted on the UC server
  - A customized module system with client branding that is hosted on the UC server
  - A customized module system with client branding that is hosted on the UC server, but that allows data sharing from the UC server to the client through specialized web services or file transfers
  - A customized module system with client branding that is hosted on the client server
  - A customized module system with client branding that is hosted on the client server and that is either (a) integrated into the existing case management system or (b) is a stand-alone system that allows information sharing with other existing systems on the client server.
- **VALIDATION SERVICES.** With the ORAS, clients retain the rights to their own data. Clients may choose to (a) conduct the validation analysis in-house, (b) send the data out to an external reviewer for validation, or (c) hire UC to perform the validation analysis.<sup>30</sup>
- **USER TRAINING.** As of this report, UC provides a 2-day basic ORAS training for \$7000, including trainer travel expenses.<sup>31</sup>

## Appendix: RNA Instrument Profile for the ORAS-CST

---

Also offered is a “train the trainer” course for those agencies that are interested in developing the internal capacity to sustain the use of the ORAS.<sup>32</sup> For more information about other training services offered by UC, contact Mr. John Schwartz at [John.Schwartz@uc.edu](mailto:John.Schwartz@uc.edu) or visit their website: <http://www.uc.edu/corrections/services/trainings.html>.

**User qualifications.** The basic user training is mandatory. This includes an overview of the ORAS tools, training on the techniques for administering and scoring individual assessments, and training on how to use the ORAS in case management.<sup>33</sup>

**Administration time.** The ORAS CST takes approximately 50 minutes to administer.<sup>34</sup>

**Modes of administration.** Information collected to complete the ORAS CST is obtained through a structured interview with the offender and an offender self-report form. Assessors are encouraged to corroborate information whenever possible with official records and collateral sources.<sup>35</sup>

**Quality assurance.** When adopting any offender assessment tool, jurisdictions must be prepared to ensure appropriate implementation and proper maintenance over time. Quality assurance recommendations and guidelines for the ORAS CST follow.<sup>36</sup>

- **OVERRIDE POLICY.** Generally, overrides may occur if (a) the user determines that the risk assessment does not reflect the actual risk of the offender and wishes to change the assessed risk level in the individual case, or (b) if, given the

assessed risk of the offender, the user must override for policy reasons (e.g., a mandate to place a particular type of offender in maximum supervision regardless of assessed risk level). ORAS developers recommend an override rate of 2-3% or less; however, overrides should not occur in more than 10% of the total population of cases and, for an individual assessor, in more than 10% of his or her caseload. If judges receive ORAS CST results, they should be notified of any override.

- **FIDELITY.** The ORAS CST interview guide is structured to increase reliability between assessors. Moreover, the automated system provided by UC includes program and data sharing features that can help minimize assessor error. However, as with any offender assessment tool, routine fidelity studies of the ORAS CST are recommended. For this purpose, the automated system includes a feature which allows the client to draw a random sample of cases (5-10%) for internal review. Clients can seek certification training from UC to learn how to conduct these studies internally.
- **INSTRUMENT REVALIDATION.** UC researchers recommend that clients revalidate the ORAS tool(s) approximately every five years.

### ENDNOTES

---

<sup>1</sup> See p. 16 in Latessa, E. J., Lemke, R., Makarios, M., Smith, P., & Lowenkamp, C. T. (2010). The creation and validation of the Ohio Risk Assessment System (ORAS)\*. *Federal Probation*, 74, 16-22. Retrieved from

## Appendix: RNA Instrument Profile for the ORAS-CST

---

[http://www.uscourts.gov/uscourts/FederalCourts/PPS/Fedprob/2010-06/02\\_creation\\_validation\\_of\\_oras.html](http://www.uscourts.gov/uscourts/FederalCourts/PPS/Fedprob/2010-06/02_creation_validation_of_oras.html)

<sup>2</sup> Brian Lovins, personal interview, February 16, 2012.

<sup>3</sup> See slide 22 in Lovins, B. (2010, September 22). *The development and implementation of the Indiana Risk Assessment System* [PowerPoint slides]. Center for Criminal Justice Research, University of Cincinnati.

<sup>4</sup> See p. 16 in Latessa et al. (2010) at endnote 1 and slide 22 in Lovins (2010) at endnote 3.

<sup>5</sup> See p. 8 in Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). *Creation and validation of the Ohio Risk Assessment System: Final report*. Cincinnati, OH: University of Cincinnati Center for Criminal Justice Research. Retrieved from [http://www.ocjs.ohio.gov/ORAS\\_FinalReport.pdf](http://www.ocjs.ohio.gov/ORAS_FinalReport.pdf)

<sup>6</sup> See p. 18 in Latessa et al. (2010) at endnote 1.

<sup>7</sup> See Latessa et al. (2010) at endnote 1.

<sup>8</sup> See Latessa, et al. (2009) at endnote 5.

<sup>9</sup> Ohio Department of Rehabilitation and Correction (2010, June 15), *Ohio Risk Assessment System*. Retrieved from <http://www.drc.ohio.gov/web/ORAS.htm>

<sup>10</sup> Counties include Ventura, Riverside and Orange in California, Alachua, Orange, and Seminole in Florida, and Dauphin, Berks, and York in Pennsylvania.

<sup>11</sup> E. Latessa, personal communication, June 11, 2014. An IRAS validation report is forthcoming. Brian Lovins, personal interview, February 16, 2012. See Goodman, M. & Thompson, L. (2011, April 13). Indiana's new risk assessment tools: What you should know. *Indiana Court Times*, 20.2, 5-7. Retrieved from

<http://issuu.com/incourts/docs/news20-2c?mode=window&backgroundColor=%02322222>. See also Lovins (2010) at endnote 3.

<sup>12</sup> B. Lovins, personal communication, February 16, 2012.

<sup>13</sup> See Latessa, et al. (2010) at endnote 1.

<sup>14</sup> See Latessa et al. (2009) at endnote 5. ORAS developers strongly recommend that these screener tools be carefully validated at the local level prior to use in other jurisdictions.

<sup>15</sup> E. Latessa, personal communication, June 11, 2014. See Latessa, E., Lovins, B., & Lux, J. (2014). *The Ohio Risk Assessment System Misdemeanor Assessment Tool (ORAS-MAT) and Misdemeanor Screening Tool (ORAS-MST)*. Cincinnati: University of Cincinnati Center for Criminal Justice Research.

<sup>16</sup> B. Lovins, personal communication, February 16, 2012.

<sup>17</sup> See Latessa et al. (2009) at endnote 5.

<sup>18</sup> See Latessa et al. (2009) at endnote 5; also B. Lovins, personal communication, February 16, 2012.

<sup>19</sup> *Id.*

<sup>20</sup> Center for Criminal Justice Research, University of Cincinnati (n.d.). *Ohio Risk Assessment System* (user manual). Cincinnati, OH: Authors.

<sup>21</sup> *Id.*

<sup>22</sup> *Id.*

<sup>23</sup> B. Lovins, personal communication, February 16, 2012.

<sup>24</sup> See Latessa et al. (2009) at endnote 5.

<sup>25</sup> See Latessa et al. (2010) at endnote 1. To assess the predictive validity of the ORAS CST, Receiver Operating Characteristics (ROC) analyses were also performed, producing Area Under the Curve (AUC) values of .71 for males and .69 for females.

<sup>26</sup> See Latessa et al. (2010) at endnote 1.

<sup>27</sup> E. Latessa, personal communication, June 11, 2014.

<sup>28</sup> E. Latessa, personal communication, June 11, 2014.

<sup>29</sup> University of Cincinnati (n.d.). *Ohio Risk Assessment System web-based modules*. Cincinnati, OH: Authors.

<sup>30</sup> B. Lovins, personal communication, February 16, 2012.

<sup>31</sup> B. Lovins, personal communication, February 16, 2012..

## Appendix: RNA Instrument Profile for the ORAS-CST

---

---

<sup>32</sup> B. Lovins, personal communication,  
December 7, 2012.

<sup>33</sup> B. Lovins, personal communication,  
December 7, 2012.

<sup>34</sup> B. Lovins, personal communication,  
December 7, 2012.

<sup>35</sup> B. Lovins, personal communication,  
December 7, 2012.

<sup>36</sup> B. Lovins, personal communication,  
December 7, 2012.

# The Static Risk and Offender Needs Guide (STRONG)

---

## STRONG GLOSSARY OF TERMS

---

<b>Risk</b>	Term is not explicitly defined in published sources on the STRONG. However, sources do refer to Andrews and Bonta's (1994) risk principle.
<b>Static risk</b>	"Risk factors that cannot decrease, such as criminal history, are static. Once a criminal record is obtained, it will always be a part of an offender's history" (Barnoski & Drake, 2007, p. 2; citing Andrews & Bonta, 1998).
<b>Dynamic risk</b>	"Dynamic risk factors, such as drug dependency, can decrease through treatment or intervention" (Barnoski & Drake, 2007, p. 2; citing Andrews & Bonta, 1998).
<b>Needs</b>	Term is not explicitly defined, but sources refer to these as dynamic, criminogenic factors that may be addressed in re-entry and supervision planning.
<b>Responsivity</b>	Term not used by STRONG developers or vendor.
<b>Protective factors</b>	Term is not explicitly defined, but sources refer to these as factors that, when present or when increased, can reduce recidivism.
<b>Strengths</b>	Term not used by STRONG developers or vendor.
<b>Recidivism</b>	The state of Washington defines recidivism as "a subsequent conviction in a Washington State Superior Court for a felony offense committed within three years of placement in the community. In addition, one year is allowed for the offense to be adjudicated in court" (Barnoski & Drake, 2007, p. 2). The static risk assessment component of the STRONG system predicts felony recidivism and distinguishes between high drug, property, and violent felony risk.

---

## HISTORY & CURRENT USE.

---

**Creation.** In 1999, the Washington State Legislature passed the Offender Accountability Act (effective July 2000), which called for improved efforts to "reduce the risk of reoffending by offenders in the community" (RCW 9.94A.010). The Washington State Institute for Public Policy

(WSIPP) was charged with evaluating the impact of these legislative changes on recidivism. In a 2003 report, WSIPP recommended improvements to the predictive accuracy of the Washington State Department of Corrections' (DOC) previous assessment tool (the LSI-R) by including more static risk items in the assessment.<sup>1</sup> The DOC requested that WSIPP create a new static risk assessment instrument comprised

## Appendix: RNA Instrument Profile for the STRONG

---

entirely of criminal history and demographic items and a new needs assessment instrument of offender deficits and protective factors for statewide use.<sup>2</sup>

WSIPP researchers developed the Static Risk Assessment in 2006 and created the Offender Needs Assessment to complete the STRONG system.<sup>3</sup> Assessments.com collaborated with a DOC team to build a software application for the STRONG and integrated it with the existing state case management system.<sup>4</sup> The STRONG was fully implemented by the Washington State DOC in August 2008.<sup>5</sup>

**Current use.** In addition to Washington State where the STRONG was developed and has been in use since 2008, the system has also reportedly been used by multiple jurisdictions in California (over 30 counties), Florida, and Texas.<sup>6</sup>

### DEVELOPMENT.

---

**Instrument purpose.** The Static Risk Assessment was designed for statewide use to assess offenders' recidivism risk, and the Offender Needs Assessment was developed to identify dynamic offender needs and protective factors that can be addressed in reentry and supervision planning.<sup>7</sup>

The Washington State DOC chose to develop static risk and offender needs assessments over the tool they previously used to assess offender risk and needs. This decision was based on the results of a WSIPP validation study on the previously used instrument. The DOC listed a number of reasons for this decision, including the increased accuracy of risk prediction in the

state with the Static Risk Assessment; greater specificity in prediction by classifying high risk offenders according to the most serious type of crime predicted (drug, property, violent); increased objectivity of a tool that is based on verifiable demographic and criminal history data rather than questions from structured interviews; decreased costs associated with the administration of the tool; and more accurate documentation of criminal history information for use in other DOC applications.<sup>8</sup>

### **Approach to instrument development.**

WSIPP researchers adopted a retrospective design in creating the Static Risk Assessment.<sup>9</sup> This means that researchers identified a "construction sample" of offenders (in this case, all 308,423 offenders released from incarceration or placed on community supervision in Washington State from 1986 to March of 2000) and used archival offender and felony reconviction data to determine which demographic and criminal history factors were most strongly associated with recidivism. Researchers applied multivariate regression techniques to identify variables that most strongly predicted recidivism for inclusion in the Static Risk Assessment tool and to develop a weighted algorithm for the calculation of risk scores. WSIPP researchers then validated this Static Risk Assessment on a sample of 51,648 Washington State felony offenders who were released from incarceration or placed on community supervision from 2001 through September 2002.

The Offender Needs Assessment was developed through a collaborative effort between WSIPP and a focus group of state correctional officers. The tool contains



## Appendix: RNA Instrument Profile for the STRONG

---

dynamic items that have a demonstrated relationship with recidivism in the broader scientific literature and also some non-criminogenic items identified by correctional officers as important for case management.<sup>10</sup> No published documentation is yet available on the development or validation of the Offender Needs Assessment tool.

### CONTENT.

---

**Structure.** The STRONG consists of two separate assessment instruments: the Static Risk Assessment and the Offender Needs Assessment. The Static Risk Assessment is designed to assess offender risk for reoffense and classify each offender to a single risk category for case management purposes. It is used to determine the amount of supervision the offender receives and the prioritization for services. Recently, Washington State conducted a study to assess the feasibility of implementing the Static Risk Assessment as a standard assessment in seven state court pretrial programs to inform pretrial release and alternative sentencing decisions. As of this report, researchers are developing a modified version of the Static Risk Assessment for use by the courts at pretrial statewide.<sup>11</sup>

The separate Offender Needs Assessment is designed to identify offender deficits and protective factors for use in guiding decisions about the type of service programming that would be most appropriate. This assessment includes dynamic criminogenic factors as well as static and non-criminogenic items identified by correctional officers as relevant to professional judgment in case planning.<sup>12</sup>

**Items and domains.** The Static Risk Assessment component of the STRONG collects information on 26 items in 6 general categories: demographic information (2 items), juvenile felony convictions and commitments (4 items), DOC commitments (1 item), felony conviction types (9 items), misdemeanor conviction types (9 items), and adult sentence violations (1 item).<sup>13</sup>

The Offender Needs Assessment component of the STRONG system in Washington State collects information on 55 items across 10 gender-neutral domains related to criminal behavior: education (4 items), community employment (10 items), friends (2 items), residential (3 items), family (8 items), alcohol and drug use (6 items), mental health (6 items), aggression (4 items), attitudes and behaviors (7 items), and coping skills (5 items). These domains assess offender needs and protective factors supported by “best practices” in the broader social learning research literature as related to criminal behavior. These factors include the presence of antisocial associates and absence of prosocial others (community employment, friends, family domains); attitudes, values, and beliefs supportive of criminal behaviors (aggression, attitudes/behaviors, coping skills domains); personality traits (alcohol/drug use, mental health, aggression domains); personal achievement (education, community employment, residential domains); and family dynamics (family domain).<sup>14</sup>

**Reporting and cutoffs.** The Static Risk Assessment groups offenders into five levels of risk (low, moderate, high drug, high property, high violent). Three separate weighted algorithms are used to compute

## Appendix: RNA Instrument Profile for the STRONG

---

general felony risk, property felony risk, and violent felony risk; these risk score calculations are used to determine the offender's classification of risk.<sup>15</sup> The Washington State DOC subsequently revised the five offender classification levels down to four groupings (low, moderate, high non-violent [property, drug], high violent).<sup>16</sup>

The Offender Needs Assessment identifies whether each of the 10 domains is considered a low, moderate, or high need and/or a low, moderate, or high protective factor. The greater the need, the more of a priority the domain is in case planning.<sup>17</sup>

### INSTRUMENT RELIABILITY AND VALIDITY.

---

**Populations studied.** The Static Risk Assessment construction and validation samples included adult community supervision and prison cohort groups in Washington State. Men and women were represented within these samples, as were various racial groups (European, African, Native, Asian, and Hispanic Americans) and types of offenses (drug, property, sex, violent non-sex offenses).<sup>18</sup> No data is yet available on the Offender Needs Assessment.

**Predictive validity.** To assess the predictive validity of the Static Risk Assessment, Receiver Operating Characteristic (ROC) analyses were performed on construction and validation samples of Washington state felony offenders on community supervision or in prison. These studies produced Area Under the Curve (AUC) values of .756 and .742 for these two samples.<sup>19</sup> No data is yet available on the validity of the Offender Needs Assessment.

**Reliability.** No data yet available.

**Potential for bias.** In the initial validation study of the Static Risk Assessment, WSIPP researchers examined the efficacy of the tool by gender and race.<sup>20</sup>

- **GENDER** The Static Risk Assessment discriminates equally well by gender (for felony offenses generally, among males, AUC = .743; among females, AUC = .720). However, Barnoski and Drake explain that the tool tends to underestimate property recidivism and overestimate violent recidivism for females compared to males.
- **RACE.** The tool also discriminates well by racial group (for felony offenses generally, among European Americans, AUC = .736; among African Americans, AUC = .723; among Native Americans, AUC = .716; among Asian Americans, AUC = .748; and among Hispanic Americans, AUC = .742). However, Barnoski and Drake explain that the tool seems to perform less well for Asian Americans in discriminating between high drug and high property recidivism.

In addition, the tool predicts violent recidivism for sex offenders but not sexual reoffending.<sup>21</sup> No data is yet available on the Offender Needs Assessment.

**Independent validation.** As of this publication, no validation studies of the STRONG have yet been published by a research organization independent from WSIPP or the Washington State DOC.

## Appendix: RNA Instrument Profile for the STRONG

---

### PRACTICAL CONSIDERATIONS.

---

**Vendor and instrument cost.** The STRONG instruments are non-proprietary. However, the STRONG software application is proprietary. Software programs for the STRONG and custom integration services are currently offered by two companies: Noble Software Group and Assessments.com. For more information on Noble Software Group, contact [info@noblesg.com](mailto:info@noblesg.com) or call (979) 248-6568. To contact an Assessments.com representative, email [info@assessments.com](mailto:info@assessments.com) or call 877-277-3778.

**Menu of other services.** Both Assessments.com and Noble Software Group offer an array of training, technical assistance, and other services to support the implementation of the STRONG.

- **IT SERVICES.** Both companies offer two general approaches to STRONG software implementation:
  - A hosted solution on remote servers for a recurring fee; may require a set-up fee
  - An enterprise solution on the agency's own servers; licensed software will run in-house, with or without customized integration.Both companies offer custom report generation and an automated case plan software product to help users build individual case plans from information on offenders' needs. Pricing is established based on the number of user licenses, not the number of assessments or reassessments.<sup>22</sup>
- **VALIDATION SERVICES.** Both companies recommend local validation of the

STRONG prior to implementation and will employ consultants to assist in this process if requested by the client. Pricing is determined based on the number of consulting hours required to conduct the validation study.<sup>23</sup>

- **USER TRAINING.** Noble Software Group and Assessments.com offer a two-day training on the STRONG, which is required before staff may use the tool. It is also strongly recommended that staff attend a two-day training on motivational interviewing before using the Offender Needs Assessment, and that users attend a booster training to enhance their skill set after they have used the STRONG for a few months.<sup>24</sup> In Washington State, probation officers in the field are trained in motivational interviewing techniques prior to conducting an Offender Needs Assessment interview.<sup>25</sup> Other training, including Train-the-Trainer programs, are also offered by both companies. As of this report, trainings typically cost approximately \$2,500.00 per day from Assessments.com.<sup>26</sup> Visit [https://www.assessments.com/content/training\\_curricula.asp](https://www.assessments.com/content/training_curricula.asp) or contact Assessments.com for more information. Trainings typically cost \$2,200 per day from Noble Software Group. Visit <http://www.noblesg.com> for more information on Noble's training programs.

**User qualifications.** The two-day STRONG training is mandatory for all users.

**Administration time.** The Static Risk Assessment component of the STRONG can

---

## Appendix: RNA Instrument Profile for the STRONG

---

take up to 15-30 minutes per offender, depending on the complexity of the offender's criminal history.<sup>27</sup> The Offender Needs Assessment takes approximately 1 hour.<sup>28</sup>

**Modes of administration.** The Static Risk Assessment is based on criminal history and demographic data extracted from case files. In Washington State, the Static Risk Assessment is conducted by a specialized, centralized unit of 13 officers<sup>29</sup> with access to out-of-state criminal history information from the Washington State Justice Information System and the National Crime Information Center. Some jurisdictions, however, opt to auto-populate the Static Risk Assessment using information from their existing management information systems.<sup>30</sup>

The Offender Needs Assessment is completed with information gathered by the probation officer from a file review, a structured interview with the offender, and collateral contacts.<sup>31</sup> Scores are automatically computed in the software application and reports are automatically generated.

**Quality assurance.** When adopting any offender assessment tool, jurisdictions must be prepared to ensure appropriate implementation and proper maintenance over time. Quality assurance recommendations and guidelines for the STRONG follow.

- **VERRIDE POLICY.** The need for an override is determined by the probation officer on a case-by-case basis and as guided by local policy. The Washington State DOC has reportedly observed a 5-10% exception rate with the tool.<sup>32</sup>
- **FIDELITY.** Assessments.com does not provide quality assurance standards for the STRONG per se. Rather, they recommend a comprehensive approach in which local implementation teams are assembled, with input from research consultants, to facilitate local decision-making about necessary business rules and continuous quality improvement needs.<sup>33</sup> Noble Software Group provides additional inter-rater reliability software products as part of a quality assurance process to ensure long-term fidelity to the instruments.  
  
In Washington State, the DOC employs trained subject matter experts who conduct routine quality assurance testing. These efforts involve observations of offender interviews and reviews of completed assessments. Additional peer support meetings and training are provided for offices struggling with quality control issues.<sup>34</sup> WSIPP developers recommend good initial training and some form of regular case review round table meetings within each unit to address quality assurance issues and to encourage ongoing dialogue about how STRONG information may be appropriately used in case management/planning.<sup>35</sup>
- **INSTRUMENT REVALIDATION.** The instrument's developer, Robert Barnoski, has indicated that the frequency of revalidation depends in part on how the instrument is used.<sup>36</sup> In Washington, the predictive accuracy of the tool is monitored annually to determine whether or not the recidivism rates

## Appendix: RNA Instrument Profile for the STRONG

---

within each risk classification level remain fairly constant. If the rates remain constant, revalidation may not be necessary. However, if evidence arises that the tool is no longer working appropriately or if significant policy changes affect the ability to use the tool as originally intended, Dr. Barnoski recommends conducting a revalidation study.

### ENDNOTES

---

<sup>1</sup> Barnoski, R., & Aos, S. (2003). *Washington's Offender Accountability Act: An analysis of the Department of Corrections' risk assessment*. Olympia, WA: Washington State Institute for Public Policy.

<sup>2</sup> Barnoski, R., & Drake, E., (2007). *Washington's Offender Accountability Act: Department of Corrections' Static Risk Instrument*. Olympia, WA: Washington State Institute for Public Policy.

<sup>3</sup> *Id.*

<sup>4</sup> Assessments.com (2011). *Washington State Department of Corrections implements the STRONG*. Bountiful, UT: Authors.

<sup>5</sup> M. Miller, personal communication, August 25, 2010.

<sup>6</sup> R. Barnoski, personal communication, April 24, 2012; also C. Lake, personal communication, August 23, 2012.

<sup>7</sup> Washington State Department of Corrections (n.d.). *Offender Needs Assessment help text*. Olympia: Washington.

<sup>8</sup> M. Miller, personal communication, August 25, 2010. See also Drake, E., & Barnoski, R. (2009). *New risk instrument for offenders improves classification decisions*. Olympia: Washington State Institute for Public Policy.

<sup>9</sup> See Barnoski & Drake (2007) at endnote 2.

<sup>10</sup> R. Barnoski, personal communication, April 24, 2012.

<sup>11</sup> Washington State Administrative Office of the Courts Information Services Division (2011). *Adult Static Risk Assessment*

feasibility study. Olympia, WA: Washington State Administrative Office of the Courts.

<sup>12</sup> R. Barnoski, personal communication, April 24, 2012.

<sup>13</sup> See Barnoski & Drake (2007) at endnote 2.

<sup>14</sup> Washington State Department of Corrections (2007). *Offender Needs Assessment tool* (factoids and weights). Olympia, WA: Authors. See also R. Barnoski, personal communication, April 24, 2012.

<sup>15</sup> See Barnoski & Drake (2007) at endnote 2 for detailed information about the classification rules and score cutoffs used to determine risk level.

<sup>16</sup> M. Miller, personal communication, August 25, 2010.

<sup>17</sup> Washington State Department of Corrections (2007) at endnote 14. See also Assessments.com (2011) at endnote 4.

<sup>18</sup> See Barnoski & Drake (2007) at endnote 2.

<sup>19</sup> *Id.*

<sup>20</sup> *Id.*

<sup>21</sup> *Id.*

<sup>22</sup> C. Lake, personal communication, August 2, 2012.

<sup>23</sup> C. Lake, personal communication, August 2, 2012.

<sup>24</sup> C. Lake, personal communication, August 2, 2012.

<sup>25</sup> M. Miller, personal communication, August 25, 2010.

<sup>26</sup> C. Lake, personal communication, August 23, 2012.

<sup>27</sup> Washington State Administrative Office of the Courts Information Services Division (2011) at endnote 11.

<sup>28</sup> R. Barnoski, personal communication, April 24, 2012.

<sup>29</sup> M. Miller, personal communication, August 25, 2010.

<sup>30</sup> R. Barnoski, personal communication, April 24, 2012.

<sup>31</sup> Washington State Department of Corrections (n.d.) at endnote 7.

<sup>32</sup> M. Miller, personal communication, August 25, 2010.

## Appendix: RNA Instrument Profile for the STRONG

---

---

<sup>33</sup> C. Lake, personal communication, August 2, 2012 and August 23, 2012.

<sup>34</sup> Assessments.com (2011) at endnote 4.

<sup>35</sup> R. Barnoski, personal communication, April 24, 2012.

<sup>36</sup> *Id.*

CENTER

FOR

COURT

INNOVATION

# Demystifying Risk Assessment

---

## Key Principles and Controversies

*Sarah Picard-Fritsche, Michael Rempel, Jennifer A. Tallon,  
Julian Adler, and Natalie Reyes*

This publication was supported by Grant No. 2011-DC-BX-K002 awarded by the Bureau of Justice Assistance to the Center for Court Innovation. The Bureau of Justice Assistance is a component of the U.S. Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the Office of Sex Offender Sentencing, Monitoring, Apprehending, Registering, and Tracking. Points of view or opinions in this document are those of the authors and do not necessarily represent the official positions or policies of the U.S. Department of Justice.

### *Acknowledgments*

The authors are indebted to Julius Lang, Director of Training and Technical Assistance with the Center for Court Innovation, for his invaluable contributions to the conceptualization of this publication. We would also like to thank Elise Jensen, Annie Schachar, Matthew Watkins, Samiha Meah and Greg Berman, also from the Center for Court Innovation, for their comments and contributions to earlier drafts of the report. Finally, we are grateful to Alissa Huntoon and her colleagues at the Bureau of Justice Assistance for their ongoing commitment to bridging the research and practice communities in the criminal justice field.



## I. Introduction

As the national push to stem the tide of mass incarceration grows, state and local jurisdictions have increasingly adopted risk assessment tools in an effort to improve decision-making at key points, such as pretrial release, sentencing, or probation and parole case management.

Today, as many as 60 risk assessment tools are in use in jurisdictions across the United States. These tools are diverse in form, length, and content. The simplest tools rely exclusively on criminal records, while others add a short defendant interview, integrating the results into a single risk score. Still other tools constitute more comprehensive risk and need assessments that require a long interview. Beyond risk classification, these longer tools offer the benefit of assessing the severity of treatable needs that are often linked to criminal behavior (“criminogenic needs”). Ultimately, diversity in the current marketplace of risk assessments should be viewed positively, as different types of tools may be more appropriate depending on the “decision point” to which they are applied (e.g., pretrial release versus correctional supervision) and the specific goals of the jurisdiction adopting the tool.

A growing body of research suggests that high quality risk assessment yields more accurate estimates of risk for future crime, when compared with professional judgment alone.<sup>1</sup> Yet despite showing strong promise for improving decision-making and mitigating the effect of cognitive biases, risk assessment tools are controversial. Specifically, debates have emerged regarding: (1) the lack of transparency of some proprietary tools; (2) the potential for risk assessment to reproduce existing racial or ethnic biases in the justice system; and (3) the inherent challenges of applying risk classifications to individual cases based on group behavior.<sup>2</sup>

Several recent articles compare the accuracy of some prominent risk assessments and propose practical criteria for tool selection,<sup>3</sup> but to date there are few, if any, pieces that address the key “big picture” questions:

1. **What is risk assessment?** How is “risk” generally defined in the field? What is data-driven risk assessment? What kinds of risk factors are commonly found in risk assessment tools and how are risk classifications created?
2. **What are some strengths and downsides?** Can risk assessment reduce unnecessary incarceration, facilitate treatment, or otherwise improve criminal justice systems? What are the limitations of current risk assessment tools and their use?
3. **Why all the debate?** What underlies current controversies regarding the use of risk assessment in criminal justice?
4. **How can the benefits of risk assessment be maximized?** What are key principles to consider for the effective, legal, and ethical application of risk assessment tools in the criminal justice field?

This essay seeks to grapple with these questions, with an eye toward bridging the worlds of research and practice. Our goal is to provide an easy-to-read overview of the

latest social science (to the extent this is possible in a field that is rapidly evolving). Our intended audience is primarily practitioners and policymakers who want to gain a better understanding of the field and have real questions about whether and how to incorporate risk assessment into their daily practice.

## II. What Is Risk Assessment?

### Defining Risk

In general, “risk” refers to the likelihood of an adverse outcome. In contemporary societies, examples of adverse outcomes include death (medicine), dropout (education), financial losses (investment), and future criminal behavior (criminal justice).

### Formal Risk Assessment

Formal risk assessment tools use large datasets regarding past trends to predict future outcomes. Risk assessment has long been entrenched in a variety of social policy arenas. What all assessments have in common is the statistical linking of likely causal factors (e.g., prior school failure) to a future outcome (e.g., likelihood of high school dropout). Despite the recent attention paid to their use in criminal justice, actuarial models are not new to this field. Statistical models that assess the relationship between criminal history, demographic factors, and re-arrest were applied to making parole decisions in Illinois as early as the 1930s.<sup>4</sup>

Within criminal justice, risk assessment has most commonly been used to predict any new criminal activity (regardless of the charge type or severity).<sup>5</sup> This definition has important limitations, especially for decision-makers at the pretrial stage who may be particularly concerned with failure-to-appear in court or risk of future violence while a current case is pending.<sup>6</sup> A robust body of scientific evidence now suggests that the likelihood of new criminal behavior can be reliably assessed based on a limited set of factors, summarized in Table 1 on pages 5-6. The table lists the most prominent predictors of recidivism risk in the left-hand column, and then presents common ways in which each factor is measured in the right-hand column.

**Table 1. Central Predictors of Recidivism Risk**

<b>Risk Factor</b>	<b>Common Measures</b>
Criminal History	Prior adult and juvenile arrests; Prior adult and juvenile convictions; Prior failures-to-appear; Other currently open cases; Prior and current charge characteristics (e.g., presence of firearms, violence, drug charges, etc.).
Demographics	Younger age; Male gender.
Antisocial Attitudes	Patterns of antisocial thinking, which typically reflect the following primary constructs: (1) Lack of empathy; (2) Externalization of blame; (3) Entitlement; (4) Attitudes supportive of violence.
Antisocial Personality Pattern	Impulsive behavior patterns; Lack of consequential thinking.
Criminal Peer Networks	Peers involved in drug use, criminal behavior and/or with a history of involvement in the justice system.
School or Work Deficits	Poor past performance in work or school (lack of a high school diploma; history of firing or suspension); Alienation from informal social control via work or school (e.g., chronic unemployment).

[continued on the next page]

**Table 1. Central Predictors of Recidivism Risk**

[continued]

Family Dysfunction	Unmarried; Recent family or intimate relationship stress; Historical lack of connection with family or intimate partner.
Substance Abuse	Duration, frequency and mode of current substance use; History of substance abuse or addiction; Self-reported drug problems.
Leisure Activities	Isolation from pro-social peers or activities.
Residential Instability	Homelessness; Frequent changes of address.
<p>Note: Factors and sample items developed based on extensive review of several comprehensive, risk-need assessment systems, including the Level of Services Inventory-Revised (1990); the COMPAS (2007); and the Ohio Risk Assessment System (2009).</p>	

As shown in Table 1, the most prominent predictors of recidivism include a mix of both “static” and “dynamic” risk factors. Static factors are those that are unchangeable either by virtue of being historical in nature (e.g., prior criminal history) or by being largely immutable characteristics of an individual (e.g., male sex). Dynamic factors are those that can be changed, such as current unemployment, substance abuse, negative peer influences, or antisocial attitudes.

The distinction between static and dynamic factors has important implications for criminal justice practice for a variety of reasons. For one, static factors—

in particular criminal history and age at arrest—are typically the strongest predictors of new criminal behavior and a short tool containing only these factors can often yield a relatively accurate risk classification. However, short static factor tools are insufficient to the larger goals of many decision-makers who are interested in reducing risk in the future. Reducing risk requires actually knowing the dynamic risk factors—does the individual in front of me have a drug problem? Are they homeless? For this reason, tools with dynamic factors tend to be more useful in contexts where it is possible to engage in risk reduction strategies (e.g., linking defendants to treatment).

### **The Theory Behind Risk Assessment**

Risk-Need-Responsivity theory was developed in the late 1980s by Canadian psychologists Don Andrews and James Bonta. A rehabilitative approach to crime prevention, this theory is grounded in research suggesting that rehabilitation, and consequently recidivism reduction, is achievable through appropriate intervention.

This theory is composed of three core principles:

- 1. The Risk Principle:**  
Risk for new criminal behavior can be predicted and that correctional interventions should focus on higher risk offenders.
- 2. The Need Principle:**  
Therapeutic interventions should be directed towards an individual's "criminogenic" needs, which are defined as dynamic needs that can be statistically tied to recidivism.<sup>7</sup>
- 3. The Responsivity Principle:**  
Correctional treatment should be adapted to the specific risk factors, needs, strengths, and other attributes of the individual.

## Risk Assessment Science and Current Practice

Social science often confirms what justice practitioners already know. For example, many prosecutors intuitively understand the importance of criminal history in predicting future offending. In other cases, however, science contradicts common assumptions. For instance, validation research in the criminal justice field has consistently shown that the presence of a diagnosis for mental illness is not a significant factor in predicting future criminal behavior, contrary to long-held assumptions in the field. Empirical research also challenges the use of current offense severity as a proxy for risk of future crime. Put simply, a felony defendant is not more likely to be re-arrested than a misdemeanor. On balance, actuarial—or data-driven—risk models have tended to outperform the judgments of individual practitioners, including clinical professionals, in accurately assessing risk. Thus the rationale behind expanding the use of formal risk assessment tools is that they offer the potential for helping justice agencies make more informed decisions.

Most assessment tools—whether they are brief tools relying exclusively on static factors or interview-based tools that include numerous risk and needs domains—are developed and tested in a similar manner. The first step is typically to start with the factors we have outlined in Table 1. The next step is to decide what additional questions might be worth asking (e.g., questions regarding perceptions of the justice system or more specific criminal background questions may be relevant depending on the context and purpose of the assessment).

Next comes testing. An empirical analysis is conducted to assess the statistical association of each selected factor on the outcome of interest (e.g., re-arrest over a certain time period). In other words, item “weights”—or the number of points assigned to each item—will be established based on the relative strength



of each risk factor in actually predicting recidivism. For example, a prior criminal conviction might be more influential than unemployment in predicting re-arrest in a test model, and will therefore be assigned a greater number of risk points in the final tool.

Finally, having weighted each factor in the tool based on its association with recidivism, risk categories will be created based upon logical “cut points” in the scoring. If the average rate of re-arrest in a sample of test cases jumps between a total score of 3 and 4, for example, this would be a logical “cut point” for a new risk category. When risk categories are accurately assigned, defendants in the higher risk groups will consistently show higher re-arrest rates.

Once a pilot version is developed, the tool is then validated. Validation simply means that the items, risk scores, and risk categories in a tool are confirmed to have a statistically significant relationship with recidivism (a statistically significant relationship is one that cannot be attributed to chance). Technically, the validation of a tool is supposed to be conducted using a fresh sample of cases, rather than the sample used to create the tool in the first place. In general, the more validation tests conducted on diverse samples of defendants, the more reliable the risk assessment tool is as a national model.

It is important to note that a validated tool is not necessarily a highly accurate tool. Predictive accuracy is typically measured by the rate at which the tool correctly classifies an individual’s risk (e.g., low, moderate, high, etc.). Any statistically validated tool will still produce false positives (individuals are predicted to re-offend but don’t) or false negatives (individuals are predicted not to re-offend but do). In simple terms, having good predictive accuracy doesn’t mean that a tool is perfect, but does mean that errors are kept relatively low.

Increasingly, tool developers are releasing Area Under the Curve (AUC) statistics, which provide a useful measure of a tool's predictive accuracy. AUC statistics range from .50 to 1.00, with a higher AUC indicating a lower rate of error in classification. By current industry standards, an AUC of .70 or higher is considered "good." An AUC in the .60 to .70 range is considered "acceptable." Given the real life consequences of criminal justice decisions, practitioners should pay close attention to AUC statistics.

### III. Can Risk Assessment Tools Improve Criminal Justice?

An individual defendant's likelihood to commit a new crime can be an important aspect of pretrial release, sentencing, community supervision, and parole decisions. Indeed, judges, prosecutors, correctional officers, and other practitioners routinely assess risk as part of their daily practice.

Because data-driven tools have been shown to improve the accuracy of risk assessments, they may improve decision-making in a variety of contexts—e.g., Is an individual a good candidate for community-based pretrial supervision? What terms of probation are appropriate in a given case? These questions, and many more, hinge on an assessment of risk. The scientific consensus is that validated risk tools with high predictive accuracy (i.e., high AUC scores) can increase the accuracy of these decisions.

In particular, risk assessment tools can help reduce recidivism by clarifying when intensive supervision or treatment is truly needed. This is a compelling justification for their use, since recidivism rates among justice-involved populations remain frustratingly high. In a national study consisting of a cohort of more than 400,000 state prisoners released in 2005, for example, 41 percent were re-arrested within a year following release.<sup>8</sup> A recent study among misdemeanor offenders in New York City serving short jail sentences has documented similarly high rates of re-arrest.<sup>9</sup> Conversely, well-implemented alternatives to incarceration such as police diversion or problem-solving courts have been shown to result in moderate, but nonetheless significant, reductions.<sup>10</sup>

However, alternatives to incarceration do not work equally well for all individuals. Meta-analyses examining over 400 studies have concluded that interventions are most effective when focused on higher-risk populations.

Indeed, intensive intervention can actually increase offending among those at lower risk.<sup>11</sup> The potential negative effects of intervention—including well-meaning treatment programs—are especially pronounced the longer and more intensive the intervention is. A recent study of one validated risk assessment tool, the LSI-R, bore this out by showing that the placement of low-risk drug court participants in long-term residential treatment doubled their likelihood of re-arrest over a two-year follow-up period.<sup>12</sup>

In sum, the literature suggests that accurate knowledge regarding criminal risk can help safely reduce the use of incarceration. A key case study that bears this out is the state of Virginia, where the use of a validated risk tool in multiple jurisdictions allowed for the diversion of 25 percent of nonviolent, prison-bound offenders over a three-year period without increasing crime.<sup>13</sup>

## IV. What Are the Limitations?

Actuarial risk assessment tools have a number of scientific and practical limitations.

### Probability, Not Perfection

No tool can predict the behavior of any individual with 100 percent accuracy. Indeed, the oft-used term “risk prediction” is misleading when applied to risk assessment tools. What these tools actually do is place individuals in a risk category (e.g., minimal, low, or high) based on the behavior of other individuals with similar characteristics. A hypothetical “high-risk” individual might have a 50 percent chance of re-arrest over a one-year period, compared with an individual in the “low-risk” category, who might have a 15 percent chance of re-arrest. These are probabilities rather than certainties. The need to tolerate some uncertainty should not come as a shock to practitioners in the criminal justice field, given the complexity of criminal behavior. At the end of the day, risk assessment is an aid—rather than a replacement—for professional discretion.

### Type of Risk

Risk assessment tools may not always be designed to assess the outcome that is most relevant to specific decision-makers. For example, a judge may be interested in risk of a new violent offense or, more specifically, a new domestic violence offense when making a pretrial release decision. Currently, many tools do not produce this type of nuance.

Additionally, only a few tools or risk assessment systems offer the ability to predict failure to appear in court, which in many jurisdictions is the most relevant question at the pretrial stage. In general, overall recidivism (any re-arrest, regardless of charge) is the easiest outcome to predict reliably. At the other end of the spectrum, failure-to-appear assessments often yield the least impressive accuracy.

## Culpability

Perhaps the least acknowledged limitation of risk assessment tools is their silence on the critical matters of moral culpability and legal proportionality. An individual's risk for re-arrest may not align intuitively with the seriousness of the current case. Individuals arrested on a low-level misdemeanor are often a high risk of re-arrest. The converse is also true; defendants charged with serious offenses may be classified as low risk. While both possibilities present challenges, the former may present a greater puzzle for the justice system. A great many defendants with relatively minor cases may be high-risk for future offending due to underlying problems like substance use, unemployment, and housing instability. A dynamic risk-needs assessment tool may aid in identifying needs, but that does not assist in crafting a sentence that is proportionate to the current offense.

## V. What Are the Major Controversies Today?

In recent days, risk assessment tools have generated a good deal of controversy, including prominent legal cases, media coverage, and even an opinion from former U.S. Attorney General Eric Holder.<sup>14</sup> What follows is a look at some of the concerns that have been raised.

### Individualized Justice

There is a legitimate concern that making risk classifications based on group behavior is a poor fit in a justice system founded on the notion of individual rights and individualized justice.<sup>15</sup> The counter-argument is that evidence-based risk assessments, and especially those assessments that measure needs as well as risk, improve the ability of the justice system to respond to each defendant's unique needs and attributes, thereby creating more just individual outcomes while protecting victims.<sup>16</sup>

### Transparency

Although there is a near consensus in the field regarding the main drivers of recidivism risk, the relative weight given to each of these factors—and the specific measures that are used—can differ significantly from one tool to the next. Often for proprietary reasons, risk assessment developers are not transparent about the weights, items, and algorithms that they are using. This lack of transparency can create a variety of challenges. Non-transparent tools may be more likely to trigger due process concerns from defendants and defense counsel.<sup>17</sup> They may also make collaborative buy-in from stakeholders regarding the use of risk assessment generally more challenging.<sup>18</sup>

## Racial Bias

There has recently been significant debate in the academic and popular press regarding the potential for actuarial risk assessment to perpetuate racial disparities, based on correlations between common risk factors (e.g., unemployment, lack of education, criminal history) and race.<sup>19</sup> Indeed, a recent study of the use of one prominent risk assessment tool in a large, urban jurisdiction, published in *ProPublica*, found that African-American defendants were more likely to be classified as high-risk for re-offense and were thus more exposed to detention when compared with white defendants.<sup>20</sup> The *ProPublica* article and a subsequent response did not resolve the more nuanced question of whether the observed race differences were due to factors external to the criminal justice system (e.g., unequal educational opportunities, employment discrimination, historic effects of neighborhood segregation) or due to racial and ethnic bias in arrest, sentencing and incarceration practices. These questions continue to be explored in the academic literature.<sup>21</sup>

Because many criminal history factors (e.g., number of prior arrests or convictions) are both correlated with race and commonly considered in sentencing decisions, there is a strong possibility that racial disparities in sentencing would persist even if there were no risk assessment tools. Indeed, risk assessment proponents argue that actuarial tools can effectively mitigate racial disproportionalities arising from implicit biases in laws, police practices, or the discretionary patterns of individual decision-makers. In Colorado, for example, an actuarial risk assessment tool effectively eliminated a pattern of disparity where judges were more likely to place African-American juveniles in secure detention compared with white juveniles with similar case characteristics.<sup>22</sup>

To date, the debate regarding race and risk assessment has been subjected to only limited rigorous



study using data from real criminal cases. An important exception is a recent study of the “PCRA,” a risk assessment tool used in federal courts, which found little to no discrepancy by race in the predictive accuracy of the tool and no significant disparate impact of the tool between black and white defendants.<sup>23</sup> These results counter the findings from ProPublica, but further research is clearly needed.

## VI. What Are Key Principles to Help the Field?

A threshold challenge for individual jurisdictions is establishing a shared understanding of the ultimate intent behind risk classification. How will the instrument be used? At what point in the process? To achieve which goals?

Answering these kinds of questions is the first step toward successful implementation. For instance, if the goal of a jurisdiction is to increase the pretrial release of low-risk individuals, the menu of appropriate assessment tools will be quite different than if the intent is to link higher-risk offenders to appropriate therapeutic intervention programs post-sentence.

In most cases, successful implementation of a formal risk assessment will require collaboration from multiple stakeholders, including judges, prosecutors, defense attorneys, and others (e.g., victim advocates and social workers). Lack of buy-in among key stakeholders has been shown to undermine the adoption of evidence-based practices more broadly, and risk-based decision making more specifically. For instance, a recent study of the use of a risk assessment system to set bail in Cook County, Illinois showed a greater than 80 percent override of the tool's recommendation on the part of arraignment court judges.<sup>24</sup> Beyond working to achieve consensus on adopting a risk-based approach, what follows are some lessons from the field about how to implement a risk assessment tool successfully.

### Reflection

Once a particular tool is adopted, the next question is how the information will be applied to decision-making. Risk assessment tools should not be thought of as a replacement for professional discretion, but rather as one of many aids to informed decision-making. Others might include legal proportionality

(i.e., the “going rates” for a particular charge) and the treatment or supervision resources in a particular jurisdiction. In short, higher risk classification suggests the need for greater resource allocation in a particular case, but this finding should be considered in context. Practitioners should use their knowledge of their reform goals, local agency culture, and target population to create guidelines for the effective application of risk assessment results. For example, if a risk-based model is adopted with the goal of creating off-ramps from incarceration for lower-risk defendants, it is incumbent on jurisdictions to identify the kinds of alternative programs that will be made available and which specific risk categories will be targeted.

### **Researcher-Practitioner Collaboration**

Given the underlying complexities of risk assessment tools and the importance of adapting risk assessments to local contexts, jurisdictions are urged to develop collaborative working groups that include both researchers and practitioners. Research-practice partnerships can enhance discussions regarding the appropriateness of specific tools. The active involvement of researchers can also facilitate local validation studies to assess predictive accuracy and racial equity of selected tools. Ongoing monitoring is key to the sustainability of risk based decision-making and provides an opportunity for jurisdictions to course correct should implementation issues arise.

Another way in which research-practice partnerships can be particularly fruitful is in the ground-up development of a risk tool specific to a certain jurisdiction or subpopulation. While tools that have been nationally tested carry the advantage of adaptability to diverse populations, localized tools are better able to account for differences in criminal risk based on geographic, social and political context. Taking a “one-size-fits-all” approach to risk assessment may

undermine successful implementation. For example, a tool validated in one jurisdiction may not be responsive to the unique risk factors and needs that are present in another jurisdiction, or a tool validated on a general pretrial population may not be responsive to the unique needs of certain defendant populations (e.g., veterans).

### **Accuracy and Transparency**

The purpose of risk assessment is simply to forecast the probability of recidivism in individual cases, with the accuracy of such predictions varying from one tool to the next, as well as from one jurisdiction to the next. If resource constraints dictate selecting a preexisting tool, practitioners are strongly encouraged to look beyond whether a tool has ever been validated and focus specifically on two performance indicators: (1) whether the type of risk assessed by the tool (re-arrest, failure to appear, new violent offense, future domestic violence) aligns with what the jurisdiction is trying to achieve; and (2) the predictive accuracy of the tool (as indicated by AUC statistics). Jurisdictions selecting preexisting tools should select one that is characterized both by strong classification accuracy and transparency. Transparency means that the weights for each risk factor in the tool are apparent to the user, as are the formulas employed to calculate the raw risk score and final risk categories. This allows jurisdictions to understand the factors driving risk in their population and supports local validation and adaptation. Conversely, proprietary risk assessment systems which only provide users with a final risk score or category will prevent this type of local control.

### **Racial Equity**

Finally, prioritizing transparency when selecting a risk assessment tool will help safeguard the assessment process from potential racial bias by allowing the jurisdiction to track disparities in risk factors, total

risk scores, and risk classifications. Detection of racial disparities in the predictive accuracy of a selected tool (i.e., different AUC statistics by race) would suggest the tool is not appropriate, while correlations between risk factors and race may suggest other empirical or policy revisions that could be made to improve implementation. For example, if unemployment status were strongly correlated with race in a particular jurisdiction, it could be removed from an assessment tool, provided it did not substantially compromise its overall predictive accuracy (empirical revision) or it might suggest the need for diversion or alternative-to-incarceration programs focused on employment needs (policy solution).

## VII. Closing

While critical debates regarding the appropriate application of actuarial models to criminal justice are likely to continue for some time, there is a growing professional consensus that the careful and ethical implementation of risk assessment tools can facilitate improved criminal justice outcomes. This paper has attempted to demystify risk-based decision-making by distilling the science underlying risk assessment and identifying some of the important benefits and limitations. Jurisdictions considering the adoption of a risk assessment tool are urged to consult the growing literature regarding the characteristics and performance of specific assessment systems and to take a localized, collaborative approach to implementation.

## Endnotes

1. Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical vs. actuarial judgment. *Science*, 243, 1668–1674; Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34, 575-607; Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98, 1060; Monahan, J., & Skeem, J. L. (2014). Risk redux: The resurgence of risk assessment in criminal sanctioning. *Federal Sentencing Reporter*, 26, 158-166.
2. Starr, S. B. (2015a). The risk assessment era: An overdue debate. *Federal Sentencing Reporter*, 27, 205–206. doi:10.1525/fsr.2015.27.4.205.
3. Desmarais, S. L., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Lexington, KY: Council of State Governments; Serin, R. & Lowenkamp, C.T. (December 2015). *Selecting and using risk and need assessments*. Alexandria, VA: National Drug Court Institute.
4. Harcourt, B. E. (2007). *Against prediction: Profiling, policing, and punishing in an actuarial age*. Chicago, IL: University of Chicago Press.
5. Most often, new criminal activity is measured as re-arrest. It is of course an imperfect measure, since not all arrests are for crimes that actually occurred, and, conversely, not all crimes that actually occurred lead to an arrest. But re-arrest is an almost universally preferred measure of re-offense for methodological reasons that lie beyond the scope of this paper to detail.
6. Some pretrial risk assessment tools (e.g., the Arnold Foundation Public Safety Assessment) include classifications along multiple dimensions of risk, including general recidivism risk, risk for violence, and risk for failure to appear. A longstanding interest in predicting new violent behavior, and in particular domestic violence, has produced quality models with some but not complete overlap with general recidivism models. A thorough review of domestic violence risk assessment can be found in a recent issue of the Domestic Violence Report: <http://www.civresearchinstitute.com/online/article.php?pid=18&iid=1210>.
7. Risk-Need-Responsivity theory has proposed all of the major factors

- included in the model of recidivism risk presented in Table 1, with the exceptions of demographic factors and residential instability. See Andrews, D., & Bonta, J. (2007). *The Risk-Need-Responsivity Model for Offender Assessment and Rehabilitation*. Public Safety Canada. Retrieved from: <https://www.publicsafety.gc.ca/cnt/rsrscs/pblctns/rsk-nd-rspnsvty/rsk-nd-rspnsvty-eng.pdf>.
8. Durose, M., Cooper, A. & Snyder, H. (2014). *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*. Washington, D.C.: Bureau of Justice Statistics. Retrieved from: <http://www.bjs.gov/content/pub/pdf/rprts05p0510.pdf>.
  9. Picard-Fritsche, S., Adler, J., Rempel, M., Reich, W. & Hynen, S. (2014). [Predictors of Re-Arrest in the Misdemeanor Population in New York City]. Unpublished raw data. New York, NY: Center for Court Innovation.
  10. Collins, S. E., Lonczak, H. S., & Clifasefi, S. L. (2015). *LEAD Program Evaluation*. Seattle, WA.: Harm Reduction Research and Treatment Lab, University of Washington: Retrieved from: [http://static1.1.sqspcdn.com/static/f/1185392/26121870/1428513375150/LEAD\\_EVALUATION\\_4-7-15.pdf](http://static1.1.sqspcdn.com/static/f/1185392/26121870/1428513375150/LEAD_EVALUATION_4-7-15.pdf). Lee, C.G., Cheesman, F., Rottman, D., Swaner, R., Lambson, S., Rempel, M. & Curtis, R. (2013). *A Community Court Grows in Brooklyn: A Comprehensive Evaluation of the Red Hook Community Justice Center*. Williamsburg VA: National Center for State Courts; Wilson, D. B., Mitchell, O., & MacKenzie, D. L. (2006). A systematic review of drug court effects on recidivism. *Journal of Experimental Criminology*, 2, 459-487.
  11. Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1997). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. In M. McShane & F. Williams (Eds.), *The Philosophy and Practice of Corrections* (pp.9-44). London, UK: Taylor and Francis.
  12. Reich, W. A., Picard-Fritsche, S., Rempel, M., & Farley, E. J. (2016). Treatment modality, failure, and re-arrest: A test of the risk principle with substance-abusing criminal defendants. *Journal of Drug Issues*, 46, 234-246.
  13. Ostrom, B. J., Kleiman, M., Cheesman, F., Hansen, R. M., & Kauder, N. B. (2002). *Offender risk assessment in Virginia*. Virginia Criminal Sentencing Commission. Retrieved from [http://www.vcsc.virginia.gov/risk\\_off\\_rpt.pdf](http://www.vcsc.virginia.gov/risk_off_rpt.pdf).



14. Holder, E. (2014). Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference. *United States Department of Justice Office of Public Affairs*. Retrieved from: <https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th>.
15. Doyle (2016, July 26). Portraits in risk. *The Crime Report*. Available at: <http://thecrimereport.org/2016/07/26/portraits-in-risk/>
16. Aldrich, L. (2016, July). The use of risk assessments in judicial decision-making. *The Domestic Violence Report*, 21, 71-72.
17. E.g., see *Wisconsin v. Loomis*, 2016 WI 68, 120 (2016).
18. Tools that are not transparent preclude stakeholder input on the ethical or legal implications of the tool's design (for example, some tools include prior arrests, while others only include prior convictions; some tools include, while others exclude, gender and age as risk factors).
19. Starr, S.B. (2015a). The risk assessment era: An overdue debate. *Federal Sentencing Reporter*, 27, 205-206; Harcourt, B.E. (2015). Risk as a proxy for race. *Federal Sentencing Reporter*, 27,237-243; Horwitz, S. (2014 August 1). Eric Holder: Basing sentences on data analysis could prove unfair to minorities. Retrieved from: [https://www.washingtonpost.com/world/national-security/us-attorney-general-eric-holder-urges-against-data-analysis-in-criminal-sentencing/2014/08/01/92d0f7ba-1990-11e4-85b6-c1451e622637\\_story.html?utm\\_term=.438a4b49cc32](https://www.washingtonpost.com/world/national-security/us-attorney-general-eric-holder-urges-against-data-analysis-in-criminal-sentencing/2014/08/01/92d0f7ba-1990-11e4-85b6-c1451e622637_story.html?utm_term=.438a4b49cc32).
20. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*. Retrieved from: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>; Dietrich, W., Mendoza, C., Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity: Performance of the COMPAS scales in Broward County. Retrieved from: [http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica\\_Commentary\\_Final\\_070616.pdf](http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf).
21. Clair, M. and Alix, S. (2016). How Judges think about racial disparities: situational decision-making in the criminal justice system. *Criminology*, 54, 332-359; Mitchell, O., & Caudy, M. S. (2015).

- Examining racial disparities in drug arrests. *JQ: Justice Quarterly*, 32, 288-313. doi:10.1080/07418825.2012.761721; Spohn, C. (2015). Race, crime, and punishment in the twentieth and twenty-first centuries. *Crime & Justice*, 44, 49-97.
22. Eaglin, J. & Solomon, D. (2016). Reducing racial & ethnic disparities in jails: Recommendations for local practice. New York: Brennan Center for Justice. Retrieved from: <https://www.brennancenter.org/sites/default/files/publications/Racial%20Disparities%20Report%20062515.pdf>.
  23. Skeem, J. & Lowenkamp, C. (2016). Risk, race & recidivism: Predictive bias & disparate impact. Retrieved from: <http://ssrn.com/abstract=2687339>.
  24. Chicago Judges Spurn Risk Assessment System in 85% of Bail Cases (2016 July 5). *The Crime Report*. Retrieved from: <http://thecrimereport.org/2016/07/05/chicago-judges-spurn-risk-assessment-system-in-85-of-bail-cases-2/>.



**Center for Court Innovation**

520 Eighth Avenue, 18th Floor

New York, New York 10018

P. 646.386.3100

F. 212.397.0985

[courtinnovation.org](http://courtinnovation.org)



**Congressional  
Research Service**

Informing the legislative debate since 1914

---

# **Risk and Needs Assessment in the Criminal Justice System**

**Nathan James**

Analyst in Crime Policy

October 13, 2015

**Congressional Research Service**

7-5700

[www.crs.gov](http://www.crs.gov)

R44087

## Summary

The number of people incarcerated in the United States has increased significantly over the past three decades from approximately 419,000 inmates in 1983 to approximately 1.5 million inmates in 2013. Concerns about both the economic and social consequences of the country's growing reliance on incarceration have led to calls for reforms to the nation's criminal justice system.

There have been legislative proposals to implement a risk and needs assessment system in federal prisons. The system would be used to place inmates in rehabilitative programs. Under the proposed system some inmates would be eligible to earn additional time credits for participating in rehabilitative programs that reduce their risk of recidivism. Such credits would allow inmates to be placed on prerelease custody earlier. The proposed system would exclude inmates convicted of certain offenses from being eligible to earn additional time credits.

Risk and needs assessment instruments typically consist of a series of items used to collect data on behaviors and attitudes that research indicates are related to the risk of recidivism. Generally, inmates are classified as being high, moderate, or low risk. Assessment instruments are comprised of static and dynamic risk factors. Static risk factors do not change, while dynamic risk factors can either change on their own or be changed through an intervention. In general, research suggests that the most commonly used assessment instruments can, with a moderate level of accuracy, predict who is at risk for violent recidivism. It also suggests that no single instrument is superior to any other when it comes to predictive validity.

The Risk-Needs-Responsivity (RNR) model has become the dominant paradigm in risk and needs assessment. The risk principle states that high-risk offenders need to be placed in programs that provide more intensive treatment and services while low-risk offenders should receive minimal or even no intervention. The need principle states that effective treatment should focus on addressing needs that contribute to criminal behavior. The responsivity principle states that rehabilitative programming should be delivered in a style and mode that is consistent with the ability and learning style of the offender.

However, the wide-scale adoption of risk and needs assessment in the criminal justice system is not without controversy. Several critiques have been raised against the use of risk and needs assessment, including that it could have discriminatory effects because some risk factors are correlated with race; that it uses group base rates for recidivism to make determinations about an individual's propensity for re-offending; and that risk and needs assessment are two distinct procedures and should be conducted separately.

There are several issues policymakers might contemplate should Congress choose to consider legislation to implement a risk and needs assessment system in federal prisons, including the following:

- Should risk and needs assessment be used in federal prisons?
- Should certain inmates be excluded from earning additional time credits?
- Should risk assessment be incorporated into sentencing?
- Should there be a decreased focus on punishing offenders?

## Contents

An Overview of Risk and Needs Assessment .....	2
Risk and Needs Factors .....	3
Can Risk and Needs Assessment Instruments Accurately Predict Risk? .....	3
How Risk and Needs Assessment is Used in the Criminal Justice System .....	4
Risk-Needs-Responsivity (RNR) Principles .....	5
Risk Principle .....	6
Needs Principle .....	6
Responsivity Principle .....	6
“Central Eight” Risk and Needs Factors .....	7
Empirical Basis for the RNR Principles .....	8
Critiques of Risk and Needs Assessment .....	9
Making Judgments about Individuals Based on Group Tendencies .....	9
Should Risk Assessment Be Separate from Needs Assessment? .....	10
Potential for Discriminatory Effects .....	10
Select Issues for Congress .....	11
Should Risk and Needs Assessment Be Used in Federal Prisons? .....	11
Should Certain Inmates Be Excluded from Earning Additional Time Credits? .....	12
Should Priority Be Given to High-Risk Offenders? .....	13
Should Risk and Needs Assessment Be Used in Sentencing? .....	13
Should There Be a Decreased Emphasis on Punishment? .....	14

## Tables

Table 1. Major Risk and Needs Factors: The “Central Eight” .....	7
Table B-1. Commonly Used Risk and Needs Assessment Instruments .....	24

## Appendixes

Appendix A. Comparison of Risk and Needs Assessment Legislation .....	17
Appendix B. Commonly Used Risk and Needs Assessment Instruments .....	23

## Contacts

Author Contact Information .....	29
----------------------------------	----

The number of people incarcerated in the United States has increased dramatically over the past three decades. In 1983, there were approximately 419,000 inmates under the jurisdiction of state and federal correctional authorities.<sup>1</sup> By the end of 2013, this figure reached approximately 1.5 million inmates.<sup>2</sup> The incarceration rate increased from 179 per 100,000 people in 1983 to 478 per 100,000 in 2013. While research indicates that the expanded use of incarceration during the 1980s and 1990s did contribute to the declining crime rate, the effect was likely small,<sup>3</sup> and incarceration has probably reached the point of diminishing returns.<sup>4</sup> Concerns about both the economic and social consequences of the country's burgeoning prison population have resulted in organizations such as Right on Crime and the Coalition for Public Safety calling for reforms to the nation's criminal justice system. Congress also formed the Charles Colson Task Force on Federal Corrections to examine the growth of the federal prison population and provide recommendations for reforms.<sup>5</sup>

There are two, not mutually exclusive, methods to reduce the number of incarcerated individuals in the United States: send fewer people to prison (e.g., placing offenders on probation or in a diversion program like a drug court) and/or release more inmates (e.g., placing inmates on parole or granting them early release by allowing them to earn more good time credits). While the ideas of diverting "low-level drug offenders" from prison or granting non-violent offenders early release so they can serve a greater proportion of their sentence in the community have been popular proposals to reduce the prison population, the crime someone is convicted of is not always the best proxy for the risk that person might pose to the community. For example, people who might not be violent individuals and who pose a low risk for future violence might be convicted of, what are legally defined as, violent crimes (e.g., illegal gun possession or driving the get-away car for someone who committed an armed robbery).<sup>6</sup> On the other hand, violent people might be sentenced to prison for non-violent crimes as a result of a plea deal.<sup>7</sup>

Because courts and correctional officials make decisions about who can safely be diverted from incarceration or granted early release, they may benefit from tools that can help in this process. Actuarial risk assessment tools may serve this purpose. Needs assessments could also help correctional officials make determinations about which offenders need higher levels of supervision and/or rehabilitative programming. Assessment instruments might help increase the efficiency of the criminal justice system by identifying low-risk offenders who could be effectively managed on probation rather than incarcerated, and they might help identify high-risk offenders who would gain the most by being placed in rehabilitative programs.

---

<sup>1</sup> University at Albany, School of Criminal Justice, Hindelang Criminal Justice Research Center, *Sourcebook of Criminal Justice Statistics (online)*, Table 6.28.2012.

<sup>2</sup> E. Ann Carson, *Prisoners in 2013*, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, NCJ 247282, Washington, DC, September 2014, p. 2.

<sup>3</sup> Research by the Brennan Center for Justice and the New York University School of Law estimates that 0%-7% of the decline in crime in the 1990s can be attributed to increased incarceration, while 0%-1% of the decrease in crime since 2000 can be attributed to increased incarceration. Oliver Roeder, Lauren-Brooke Eisen, and Julia Bowling, *What Caused the Crime Decline?*, Brennan Center for Justice, New York, NY, February 12, 2015, p. 6.

<sup>4</sup> Anne Morrison Piehl and Bert Useem, "Prisons," in *Crime and Public Policy*, ed. Joan Petersilia and James Q. Wilson, 2<sup>nd</sup> ed. (New York: Oxford University Press, 2011), p. 542.

<sup>5</sup> See P.L. 113-76 and the joint explanatory statement to accompany P.L. 113-76, printed in the January 15, 2014, *Congressional Record*, p. H514.

<sup>6</sup> Leon Neyfakh, "OK, So Who Gets to Go Free?" *Slate*, March 4, 2015, [http://www.slate.com/articles/news\\_and\\_politics/crime/2015/03/prison\\_reform\\_releasing\\_only\\_nonviolent\\_offenders\\_won\\_t\\_get\\_you\\_very\\_far.html](http://www.slate.com/articles/news_and_politics/crime/2015/03/prison_reform_releasing_only_nonviolent_offenders_won_t_get_you_very_far.html).

<sup>7</sup> *Ibid.*



The use of risk and needs assessment in the criminal justice system is not without controversy, however. Proponents of assessment assert that the tools used to assess the risk and needs of inmates are better than the independent judgment of clinicians and that the tools have demonstrated the ability to make distinctions between high- and low-risk offenders.<sup>8</sup> Nonetheless, risk and needs assessment is not 100% accurate. Two experts in the field note that “[a]lthough statistical risk assessment reduces uncertainty about an offender’s probable future conduct, it is subject to errors and should be regarded as advisory rather than peremptory. Even with large data sets and advanced analytical techniques, the best models are usually able to predict recidivism with about 70% accuracy—provided it is completed by trained staff.”<sup>9</sup>

There have been legislative proposals introduced in the current Congress that would require the Bureau of Prisons (BOP) to implement a risk and needs assessment system.<sup>10</sup> The system would evaluate inmates and place inmates in rehabilitative programs and productive activities. Under the proposed system some inmates would be allowed to earn additional time credits for participating in rehabilitative programs that reduce their risk of recidivism. Such credits would allow inmates to be placed in prerelease custody earlier.

This report provides information on the use of risk and needs assessment in the criminal justice system. It starts with an overview of risk and needs assessment and a discussion of some of the critiques of it. The report concludes with a discussion of the issues policymakers might consider if they debate legislation to expand the use of risk and needs assessment in the federal prison system.

## An Overview of Risk and Needs Assessment

A risk and needs assessment instrument measures offenders’ criminal risk factors and specific needs that if addressed will reduce the likelihood of future criminal activity.<sup>11</sup> Assessment instruments typically consist of a series of questions that help guide an interview with an offender in order to collect data on behaviors and attitudes that research indicates are related to the risk of recidivism.<sup>12</sup> Data collected during the interview is typically supplemented with information from an official records check, such as a criminal history records check.<sup>13</sup> A total score is calculated using the risk and needs assessment instrument, and that score places the offender into a risk category (typically “low,” “moderate,” or “high”).

---

<sup>8</sup> Eileen Sullivan and Ronnie Green, “States Predict Inmates’ Future Crimes with Secretive Surveys,” *Associated Press*, February 24, 2015, <http://bigstory.ap.org/article/027a00d70782476eb7cd07fbcca40fc2/states-predict-inmates-future-crimes-secretive-surveys>.

<sup>9</sup> Edward J. Latessa and Brian Lovins, “The Role of Offender Risk Assessment: a Policy Maker Guide,” *Victims and Offenders*, vol. 5, 2010, p. 212 (hereinafter “The Role of Offender Risk Assessment”).

<sup>10</sup> See for example, S. 467, S. 2123, H.R. 759, and H.R. 2944. A more detailed comparison of the four bills can be found in **Appendix A**.

<sup>11</sup> Pew Center on the States, *Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders*, issue brief, September 2011, [http://www.pewtrusts.org/~media/legacy/uploadedfiles/pcs\\_assets/2011/PewRiskAssessmentbriefpdf.pdf](http://www.pewtrusts.org/~media/legacy/uploadedfiles/pcs_assets/2011/PewRiskAssessmentbriefpdf.pdf), p. 2 (hereinafter, “*Risk/Needs Assessment 101*”).

<sup>12</sup> *Ibid.*

<sup>13</sup> *Ibid.*

## Risk and Needs Factors

Generally speaking, risk and needs assessment instruments typically consist of both static and dynamic risk factors. Static risk factors do not change over time. Examples include age at first arrest, gender, past problems with substance or alcohol abuse, prior mental health problems, or a past history of violating terms of supervision (e.g., parole or probation).<sup>14</sup>

Dynamic risk factors, also called “criminogenic<sup>15</sup> needs,” change and/or can be addressed through interventions. Examples include current age, education level, or marital status; being currently employed or in substance or alcohol abuse treatment; and having a stable residence.<sup>16</sup>

## Can Risk and Needs Assessment Instruments Accurately Predict Risk?

In general, research indicates that most commonly used risk and needs assessment instruments can, with a moderate level of accuracy, predict who is at risk for recidivism.<sup>17</sup> It also indicates that no one instrument is superior to any other when it comes to predictive validity.<sup>18</sup> One group of researchers concluded that “[o]verall, our results showed that all of the nine tools predicted violence at above-chance levels, with medium effect sizes, and no one tool predicting violence significantly better than any other. In sum, all did well, but none came first.”<sup>19</sup>

The relative interchangeability of risk and needs assessment instruments was demonstrated by an experiment whereby items from four instruments were written on pieces of paper and placed in a coffee can, and researchers drew 13 of the items from the coffee can at random to create four new instruments. The researchers found that the four “coffee can” instruments predicted violent recidivism as well as the four original needs and risk assessment instruments.<sup>20</sup>

Two scholars have posited that there might be two explanations for why well-validated risk and needs assessment instruments have similar levels of performance. First, some evidence suggests that there is a “natural limit” to the predictive utility of instruments.<sup>21</sup> Simply stated, there is a limit to how accurately recidivism can be predicted given society’s current level of knowledge about criminal behavior. Second, well-validated instruments may show similar levels of performance because they are tapping “common factors” or shared dimensions of risk, even

---

<sup>14</sup> James Austin, “The Proper and Improper Use of Risk Assessment in Corrections,” *Federal Sentencing Reporter*, vol. 16, no. 3, February 2004, p. 5 (hereinafter “The Proper and Improper Use of Risk Assessment in Corrections”).

<sup>15</sup> “Criminogenic” is commonly understood to mean factors that can contribute to criminal behavior.

<sup>16</sup> “The Proper and Improper Use of Risk Assessment in Corrections,” p. 5.

<sup>17</sup> **Appendix B** provides information on some commonly used risk and needs assessment instruments.

<sup>18</sup> Public Safety Canada, *Predicting Violent Recidivism*, Research Summary, vol. 12, no. 3, May 2007, <http://www.publicsafety.gc.ca/cnt/rsrscs/pblctns/prdng-rcvds/index-eng.aspx>; Mary Ann Campbell, Sheila French, and Paul Gendreau, “The Prediction of Violence in Adult Offenders; A Meta-Analytic Comparison of Instruments and Methods of Assessment,” *Criminal Justice and Behavior*, vol. 36, no. 6, June 2009, pp. 567-590; Min Yang, Stephen C.P. Wong, and Jeremy Coid, “The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools,” *Psychological Bulletin*, vol. 136, no. 5, 2010, pp. 740-767.

<sup>19</sup> Min Yang, Stephen C.P. Wong, and Jeremy Coid, “The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools,” *Psychological Bulletin*, vol. 136, no. 5, 2010, p. 757.

<sup>20</sup> Daryl G. Kroner, Jeremy F. Mills, and John R. Reddon, “A Coffee Can, Factor Analysis, and Prediction of Antisocial Behavior: The Structure of Criminal Risk,” *International Journal of Law and Psychology*, vol. 28, 2005, pp. 360-374.

<sup>21</sup> John Monahan and Jennifer L. Skeem, “Risk Redux: The Resurgence of Risk Assessment in Criminal Sanctioning,” *Federal Sentencing Reporter*, vol. 26, no. 3, February 2014, p. 162 (hereinafter “Risk Redux”).

though the instruments utilize different items or have different approaches.<sup>22</sup> For example, the researchers who conducted the “coffee can” experiment found that assessment instruments gauge four overlapping dimensions: criminal history, persistent antisocial lifestyle, psychopathic personality, and alcohol/mental health issues.

## How Risk and Needs Assessment is Used in the Criminal Justice System

Risk and needs assessment can be used at nearly all points of the criminal justice system, as highlighted by a Vera Institute of Justice memorandum:<sup>23</sup>

- **Pretrial detention:** Courts use risk assessment instruments to help them make decisions about which defendants can be safely released pending trial. The assessment typically measures the likelihood the defendant will appear if released and whether the defendant is likely to commit another offense while on release.
- **Sentencing:** Risk and needs assessment can be used to help a sentencing judge decide whether an offender should be incarcerated or placed on community supervision. The result of the assessment can also help the judge decide whether any conditions should be placed on the offender.
- **Probation/Post-Release Supervision:** Probation and parole agents use risk and needs assessment instruments to predict the likelihood that offenders will recidivate and to identify offenders’ criminogenic needs. The results of the assessment help probation and parole agents make decisions about (1) the level of supervision offenders will receive, (2) developing an individualized case management plan that focuses on placing offenders in programs that help reduce their risk of recidivism; and (3) sanctions for violations of the conditions of release.
- **Prison:** Correctional authorities use risk assessment to make decisions about the security level to which inmates will be assigned (e.g., a high, medium, low, or minimum security facility). Prison classification systems traditionally try to identify inmates who are at a high risk for escaping or who might be management problems.
- **Parole Boards and Releasing Authorities:** Risk assessment can be used by parole boards and releasing authorities to make decisions about which inmates can be safely released from incarceration.

Two experts on the use of risk and needs assessment note that while there is evidence that risk and needs assessment is widely used in corrections, there is a great deal of variation in how it is implemented and employed.<sup>24</sup> Some states have adopted and implemented standardized assessment instruments that are used throughout the state and across a wide variety of settings.<sup>25</sup> Other states use risk and needs assessment in a less systematic manner. Ohio is highlighted as a

---

<sup>22</sup> Ibid.

<sup>23</sup> Vera Institute of Justice’s Center of Sentencing and Corrections, *Risk and Needs Assessment*, memorandum to the Delaware Justice Reinvestment Task Force, October 12, 2011, pp. 9-12, [https://ltgov.delaware.gov/taskforces/djrtf/DJRTF\\_Risk\\_Assessment\\_Memo.pdf](https://ltgov.delaware.gov/taskforces/djrtf/DJRTF_Risk_Assessment_Memo.pdf) (hereinafter “Vera Institute of Justice’s memorandum re: risk and needs assessment”).

<sup>24</sup> The Role of Offender Risk Assessment, p. 205.

<sup>25</sup> Ibid.

noteworthy example because the state developed a statewide risk and needs assessment system that is used across all levels of its correctional system.

### **An Example of Risk and Needs Assessment from Ohio**

The Ohio Risk Assessment System (ORAS) provides an example of how risk and needs assessment can be integrated into the criminal justice system. Ohio passed a law that required the Ohio Department of Rehabilitation and Corrections to develop a risk assessment tool to evaluate the likelihood of recidivism for adult offenders.<sup>26</sup> The law required the risk assessment tool to be used by

- each municipal, county, and common pleas court, when it orders an assessment for sentencing or other purposes,
- the probation department serving those courts,
- state and local correctional institutions,
- private correctional institutions,
- community-based correctional facilities, and
- the Adult Parole Authority and the Ohio Parole Board.

ORAS was “developed as a statewide system to assess the risk and needs of Ohio offenders in order to improve consistency and facilitate communication across criminal justice agencies.”<sup>27</sup> The goal was to develop assessment tools that were predictive of recidivism at different stages in the criminal justice system; specifically, pretrial release, community supervision, prison intake, and community reentry. The ORAS consists of seven different tools that are used at various points in the criminal justice system:

- the Pre-Trial Tool (PAT),
- the Community Supervision Screening Tool (CSST),
- the Community Supervision Tool (CST),
- the Prison Screening Tool (PST),
- the Prison Intake Tool (PIT),
- the Reentry Tool (RT), and
- the Supplemental Reentry Tool (SRT)

## **Risk-Needs-Responsivity (RNR) Principles**

The Risk-Needs-Responsivity (RNR) model of risk and needs assessment and offender treatment incorporates many of the evidence-based practices for reducing recidivism.<sup>28</sup> As the name implies, the model has three main principles: assessing risk, addressing criminogenic needs, and providing treatment that is responsive to the offender’s abilities and learning style.<sup>29</sup>

<sup>26</sup> Ohio Department of Rehabilitation and Corrections, *Ohio Risk Assessment System*, <http://www.drc.ohio.gov/web/oras.htm>.

<sup>27</sup> Edward J. Latessa, Richard Lemke, and Matthew Makarios, et al., “The Creation and Validation of the Ohio Risk Assessment System (ORAS),” *Federal Probation*, vol. 74, no. 1 (June 2010).

<sup>28</sup> Pamela M Casey, Roger K. Warren, and Jennifer K. Elek, *Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group*, National Center for State Courts, Williamsburg, VA, 2011, p. 5.

<sup>29</sup> There are several principles other than risk, needs, and responsivity that are a part of the RNR model. These include three overarching principles: delivering services with respect for people, basing programs on psychological theory, and reducing criminal victimization. In addition to the risk, needs, and responsivity principles, there are several other core principles, including introducing human services in order to reduce recidivism, targeting more criminogenic needs relative to noncriminogenic needs, assessing offenders’ strengths to enhance prediction and specific responsivity effects, using structured assessments, and only using professional discretion for very specific reasons. There are also three organizational principles: a preference for community-based services, services are enhanced when delivered by (continued...)

The RNR model is based on the social psychology of offending, which posits that individuals and social/situational factors intersect to create values, cognitions, and personality orientations that are conducive to criminal conduct.<sup>30</sup> These ways of thinking and responding are learned and become reinforced through feedback, and they eventually result in individual differences in the propensity for criminal behavior.<sup>31</sup> The RNR model has become a dominant paradigm in the assessment literature because it is one of the few comprehensive theories of how to provide effective intervention to offenders. Experts in the field of risk and needs assessment assert that assessment systems should adhere to the RNR model. As the Vera Institute of Justice notes, “[u]nderlying the development of evidence-based practices in the criminal justice system are the *risk, need, and responsivity* principles” [emphasis original].<sup>32</sup>

Many other theories of criminal behavior focus on the social causes of criminal behavior, factors that cannot be addressed through treatment. On the other hand, the RNR model focuses on the proximate causes of criminal behavior, which can be the focus of effective correctional treatment.

## Risk Principle

The risk principle has two aspects: (1) criminal behavior can be predicted, and (2) the level of treatment should be matched to the risk level of the offender.<sup>33</sup> The risk principle states that high-risk offenders need to be placed in programs that provide more intensive treatment and services while low-risk offenders should receive minimal or even no intervention.

## Needs Principle

The needs principle states that effective treatment should focus on addressing criminogenic needs, that is, dynamic risk factors that are highly correlated with criminal conduct.<sup>34</sup> Also, according to the needs principle, effective treatment should not focus on addressing noncriminogenic needs, because changes in noncriminogenic needs are not associated with reduced recidivism.<sup>35</sup>

## Responsivity Principle

The responsivity principle states that rehabilitative programming should be delivered in a style and mode that is consistent with the ability and learning style of the offender.<sup>36</sup> The responsivity principle is further divided into two elements. The general responsivity principle states that

---

(...continued)

therapists and staff with high-quality relationships skills in combination with high-quality structuring skills, and management should closely oversee the provision of services. For a more detailed overview of all of the principles of the RNR model, see Chapter 2 of D.A. Andrews and James Bonta, *The Psychology of Criminal Conduct*, 5<sup>th</sup> ed. (New Providence, NJ: Anderson Publishing, 2010) (hereinafter “*The Psychology of Criminal Conduct*”).

<sup>30</sup> Francis T. Cullen and Cheryl Lero Jonson, “Rehabilitation and Treatment Programs,” in *Crime and Public Policy*, ed. James Q. Wilson and Joan Petersilia, 2<sup>nd</sup> ed. (New York: Oxford University Press, 2011), p. 319.

<sup>31</sup> Ibid.

<sup>32</sup> Vera Institute of Justice’s memorandum re: risk and needs assessment, p. 2.

<sup>33</sup> *The Psychology of Criminal Conduct*, p. 47.

<sup>34</sup> The Role of Offender Risk Assessment, p. 209.

<sup>35</sup> *The Psychology of Criminal Conduct*, p. 49.

<sup>36</sup> Ibid., p. 49.

cognitive-behavioral and social learning therapies are the most effective form of intervention.<sup>37</sup> The specific responsivity principle states that treatment should consider the relevant characteristics of the offender (e.g., the offender’s motivations, preferences, personality, age, gender, ethnicity, and cultural identification, along with other factors).

### “Central Eight” Risk and Needs Factors

The developers of the RNR principles identified what they deem the “central eight” risk and needs factors. These risk and needs factors include the “big four,” which they believe to be the “major predictor variables and indeed the major causal variable in the analysis of criminal behavior in individuals.”<sup>38</sup> The remaining four risk and needs factors are referred to as the “moderate four.” The “central eight” risk and needs factors are presented in **Table 1**.

Even though antisocial behavior is the most prominent of the “central eight” risk and needs factors, a common mistake in risk assessment is conflating past antisocial behavior with current antisocial behavior. The seriousness of the current offense is *not* a risk factor.<sup>39</sup> A past history of antisocial behavior is what indicates a risk of future offending.

**Table 1. Major Risk and Needs Factors: The “Central Eight”**

Risk/Need Factor	Indicator	Target for Intervention
<b>The Big Four</b>		
History of Antisocial Behavior	This includes early involvement in any number of a variety of antisocial activities. Major indicators include being arrested at a young age, a large number of prior offenses, and rule violations while on conditional release.	History cannot be changed, but targets for change include developing new noncriminal behaviors in high-risk situations and building self-efficacy beliefs supportive of reform.
Antisocial Personality Pattern	People with this factor are impulsive, adventurous, pleasure-seeking, involved in generalized trouble, restlessly aggressive, and show a callous disregard for others.	Building skills to address weak self-control, anger management, and poor problem-solving.
Antisocial Cognition	People with this factor hold attitudes, beliefs, values, rationalizations, and personal identity that is favorable to crime. Specific indicators include identifying with criminals, negative attitudes towards the law and justice system, beliefs that crime will yield rewards, and rationalizations that justify criminal behavior (e.g., the “victim deserved it”).	Reducing antisocial thinking and feelings through building and practicing less risky thoughts and feelings.
Antisocial Associates	This factor includes both association with procriminal others and isolations from anticriminal others.	Reduce association with procriminal others and increase association with anticriminal others.

<sup>37</sup> Ibid., pp. 49-50.

<sup>38</sup> Ibid., p. 55.

<sup>39</sup> Ibid., p. 60.

Risk/Need Factor	Indicator	Target for Intervention
<b>The Moderate Four</b>		
Family/Marital Circumstances	Poor-quality relationships between either the child and the parent (in the case of juvenile offenders) or spouses (in the case of adult offenders) in combination with either neutral expectations with regards to crime or procriminal expectations.	Reduce conflict, build positive relationships, and enhance monitoring and supervision.
School/Work	Low levels of performance and involvement and low levels of rewards and satisfaction.	Enhance performance, involvement, rewards, and satisfaction.
Leisure/Recreation	Low levels of involvement in and satisfaction from noncriminal leisure pursuits.	Enhance involvement in and satisfaction from noncriminal leisure activities.
Substance Abuse	Problems with abusing alcohol and/or other drugs (excluding tobacco). Current problems with substance abuse indicate a higher risk than past substance abuse problems.	Reduce substance abuse, reduce the personal and interpersonal supports for substance-oriented behavior, and enhance alternatives to substance abuse.

**Source:** Adapted from Table 2.5 in D.A. Andrews and James Bonta, *The Psychology of Criminal Conduct*, 5<sup>th</sup> ed. (New Providence, NJ: Anderson Publishing, 2010).

## Empirical Basis for the RNR Principles

Research on the risk principle suggests that recidivism is only reduced when high-risk offenders are placed in programs where they receive intensive levels of services.<sup>40</sup> In some instances, research also found that low-risk offenders who were placed in intensive treatment programs actually had an increased likelihood of recidivism.<sup>41</sup> This could be because placing low-risk offenders in intensive programming interrupts support structures or self-correcting behaviors that already exist, or because it exposes low-risk offenders to high-risk offenders who may have a negative influence on low-risk offenders' thoughts or behaviors.<sup>42</sup>

Research suggests that programs that adhere to the RNR principles are more effective at reducing recidivism.<sup>43</sup> Specifically, the more of the RNR principles a treatment program adheres to, the greater the reduction in recidivism. Research also indicates that treatment can be more effective when provided in a community setting, though treatment that adheres to the RNR principles can still be effective when provided in a custodial setting (i.e., prison or jail).

The developers of the RNR principles argue that research results indicate that the “central eight” risk and needs factors are the best predictors of future criminal behavior. A review of eight meta-analyses on the relationship between certain risk and needs factors and criminal behavior found

<sup>40</sup> Ibid., p. 48.

<sup>41</sup> Ibid.

<sup>42</sup> Vera Institute of Justice's memorandum re: risk and needs assessment, p. 2.

<sup>43</sup> James Bonta and D.A. Andrews, *Risk–Need–Responsivity Model for Offender Assessment and Rehabilitation*, Public Safety Canada, June 2007, pp. 9-12, <http://www.publicsafety.gc.ca/cnt/rsrscs/pblctns/rsk-nd-rspnsvty/index-eng.aspx>, hereinafter, “*RNR Model for Offender Assessment and Rehabilitation*.”

moderate effect sizes for both the “big four” and the “moderate four” risk factors. In comparison, the mean effect size for four minor risk factors was not statistically significant.

## Critiques of Risk and Needs Assessment

Proponents assert that risk and needs assessment instruments are effective enough that they can help officials make decisions about who needs to be incarcerated and who can be safely treated and supervised in the community. However, while risk and needs assessment instruments have demonstrated the ability to predict the risk of recidivism with some degree of accuracy, there are people who are concerned about how these instruments are used in the criminal justice system. One expert notes that risk and needs assessment involves judgments about uncertainty.<sup>44</sup> Risk and needs assessment can limit the range of plausible speculation about a potential outcome, but it will never be certain. This expert notes that there are so many determinants of human behavior that it is impossible to reason through all of the possible outcomes. This section of the report provides an overview of some of the critiques of risk and needs assessment.

### Making Judgments about Individuals Based on Group Tendencies

One of the key critiques of risk and needs assessment is that while there is evidence of some predictability in group behavior, it is difficult, if not impossible, to make a determination about how individual members of a group will behave.<sup>45</sup>

Two scholars argue that “[o]n the basis of empirical finding, statistical theory, and logic, we conclude that predictions of future offending [using risk and needs assessment] cannot be achieved in the individual case with any degree of confidence.”<sup>46</sup> They note that it is a logical fallacy to make a causal inference about a member of a group based on the group’s characteristics.<sup>47</sup>

However, the supposition that risk and needs assessment provides no useful information for criminal justice decision making has been vigorously contested. Two scholars assert that while the probabilities associated with assessment clearly will never be certain, group data can help criminal justice professionals make decisions about who is at risk of recidivating.<sup>48</sup> Proponents of the use of assessment note that the insurance industry makes decisions about risk based on actuarial methods.<sup>49</sup> Insurance companies set the price for insurance on a purchaser’s membership in a group. Without relying on such probabilities it would be impossible for insurance companies to set prices.

However, researchers who question the use of risk and needs assessment to predict individual risk assert that this analogy is false because insurance companies are interested in predicting what

---

<sup>44</sup> R. Karl Hanson, “The Psychological Assessment of Risk for Crime and Violence,” *Canadian Psychology*, vol. 50, no. 3 (2009), p. 172.

<sup>45</sup> James Austin, “The Proper and Improper Use of Risk Assessment in Corrections,” *Federal Sentencing Reporter*, vol. 16, no. 3, February 2004, p. 3 (hereinafter, “The Proper and Improper Use of Risk Assessment in Corrections”).

<sup>46</sup> David J. Cook and Christine Michie, “Limitations of Diagnostic Precision and Predictive Utility in the Individual Case: A Challenge for Forensic Practice,” *Law and Human Behavior*, vol. 34, 2010, p. 259 (hereinafter, “Limitations of Diagnostic Precision and Predictive Utility in the Individual Case”).

<sup>47</sup> *Ibid.*, p. 271.

<sup>48</sup> Jennifer L. Skeem and John Monahan, *Current Directions in Violence Risk Assessment*, University of Virginia Law School, Public Law and Legal Theory Research Paper Series, no. 2011-13, March 2011, pp. 8-9.

<sup>49</sup> *Ibid.*



proportion of insured individuals will, for example, die within a certain time frame; they are not interested in predicting the deaths of certain individuals.<sup>50</sup>

## Should Risk Assessment Be Separate from Needs Assessment?

Research suggests that including dynamic risk factors in risk and needs assessment can increase its accuracy.<sup>51</sup> However, some experts in the field have also advocated for shorter risk assessment instruments that focus on a relatively short list of static risk factors.

One scholar of risk and needs assessment argues that risk and needs should not be measured together. He notes that many early assessment instruments were simple and consisted of fewer than a dozen factors.<sup>52</sup> More recently the focus of risk assessment has changed from solely predicting risk to “risk reduction.” The focus on risk reduction means that instruments added dynamic risk factors that can change with time and/or are amenable to treatment and, therefore, reduce the offender’s risk level.<sup>53</sup> However, some research has shown that *some* dynamic risk factors are not related to any measure of recidivism.<sup>54</sup> Also, dynamic risk factors might be more difficult to measure accurately.<sup>55</sup>

It is argued that the inclusion of a bevy of dynamic risk factors has diluted the ability of risk and needs assessment instruments to classify cases accurately.<sup>56</sup> Most assessment instruments, even though they contain risk factors that might be extraneous to predicting risk, contain enough valid risk factors that they are able to predict with modest accuracy which groups of offenders are the most likely to recidivate. However, “[t]here is substantial evidence available to suggest that relatively brief risk indices outperform longer, more complex models.”<sup>57</sup> For example, one study in Pennsylvania found that risk assessment accuracy was improved by using only 8 of the 54 factors in one commonly used instrument.

Two scholars have argued that risk assessment should be conducted separately from needs assessment.<sup>58</sup> Combining risk and needs assessment has the potential to introduce variables that might be useful when trying to assess what interventions would be effective to *reduce* an offender’s risk, but it might reduce the ability of the instrument to predict risk accurately in situations where *only* predicting risk is all that is warranted (e.g., should someone be granted pretrial release or should an inmate be released on parole).

## Potential for Discriminatory Effects

There is a concern that the wide-scale use of risk and needs assessment might exacerbate racial disparities in the nation’s prison systems. One scholar contends that research on assessment

---

<sup>50</sup> Limitations of Diagnostic Precision and Predictive Utility in the Individual Case, p. 271.

<sup>51</sup> Stephen D. Gottfredson and Laura J. Moriarty, “Statistical Risk Assessment: Old Problems and New Applications,” *Crime and Delinquency*, vol. 52, no. 1, January 2006, p. 191 (hereinafter “Statistical Risk Assessment: Old Problems and New Applications”).

<sup>52</sup> Christopher Baird, *A Question of Evidence: A Critique of Risk Assessment Models Used in the Justice System*, National Council of Crime and Delinquency, February 2009, p. 3 (hereinafter “*A Question of Evidence*”).

<sup>53</sup> *Ibid.*

<sup>54</sup> *Ibid.*

<sup>55</sup> Statistical Risk Assessment: Old Problems and New Applications, p. 191.

<sup>56</sup> *A Question of Evidence*, p. 3.

<sup>57</sup> *Ibid.*, p. 5.

<sup>58</sup> Statistical Risk Assessment: Old Problems and New Applications, p. 192.

instruments has not adequately vetted the tools for use on racial minorities.<sup>59</sup> This scholar notes that social context, such as gender, race, and economic and socio-structural factors, plays a role in crime, and assessment does not account for these factors.<sup>60</sup>

It is also possible that minorities might score higher on risk and needs assessments because “of their elevated exposure to risk, racial discrimination, and social inequality—not necessarily because of their criminal propensities or the crimes perpetrated.”<sup>61</sup> One expert noted that most instruments use socioeconomic factors that correlate with race and ethnicity, and include factors that punish people for choices that people are allowed to make in a free society (e.g., whether to get married, live in a stable residence, or have a regular job).<sup>62</sup>

Another researcher has warned of the need to thoroughly evaluate risk and needs assessment instruments to ensure that the classifications of risk are not biased against African-Americans and Hispanics.<sup>63</sup> Cutoff points developed using reoffending rates for white offenders might lead to over- or under-classification for some minorities.

A review of the research on the relationship between race/ethnicity and predictive validity of risk and needs assessment found contradictory and mixed results.<sup>64</sup> The researchers found a total of eight meta-analyses that evaluated the role that race/ethnicity played in mediating the ability of instruments to predict recidivism. Three studies found that the higher the percentage of white offenders in the sample, the higher the predictive validity of the instrument—suggesting that instruments can better predict risk for white offenders. The other five studies found no evidence that predictive validity varied based on the race/ethnicity of the participants.

## Select Issues for Congress

There are four pieces of legislation before Congress that would establish a risk and needs assessment system in the BOP. The above discussion about the strengths and weaknesses of assessment might raise a question among some policymakers about whether the BOP should use a risk and needs assessment system. Even if policymakers decide that the BOP should use assessment, there might be additional questions about how to implement an effective assessment system. The four legislative proposals might also raise questions about whether other measures should be taken in order to reduce the number of inmates in federal prisons. This section of the report discusses some of the issues that might arise if Congress considers any of the current legislative proposals.

### Should Risk and Needs Assessment Be Used in Federal Prisons?

An overarching issue policymakers might consider is whether the BOP should use risk and needs assessment. Research suggests that assessment instruments can make distinctions between high-

---

<sup>59</sup> Kelly Hannah-Moffat, *Actuarial Sentencing: An “Unsettled” Proposition*, paper presented at University at Albany Symposium on Sentencing, September 2010, p. 16 (hereinafter “*Actuarial Sentencing: An ‘Unsettled’ Proposition*”).

<sup>60</sup> *Ibid.*, p. 14.

<sup>61</sup> *Ibid.*, p. 17.

<sup>62</sup> Michael Tonry, “Legal and Ethical Issues in the Prediction of Recidivism,” *Federal Sentencing Reporter*, vol. 26, no. 3, February 2014, p. 171.

<sup>63</sup> *The Psychology of Criminal Conduct*, p. 333.

<sup>64</sup> Jay P. Singh and Seena Fazel, “Forensic Risk Assessment: a Metareview,” *Criminal Justice and Behavior*, vol. 37, no. 9, September 2010, p.978.

and low-risk offenders with some degree of accuracy. Furthermore, assessment systems that adhere to the RNR principle appear to be effective at reducing recidivism. Implementing an assessment system in federal prisons would appear, based on the current research, to be an evidence-based way to improve the effectiveness of rehabilitative programming, and when combined with additional time credits for some inmates who participate in rehabilitative programs and productive activities, it might provide a means for reducing the federal prison population without increasing the risk to public safety.

However, risk and needs assessment systems are not flawless. There will always be false positives (e.g., inmates who are determined to be high risk but are actually a low risk for recidivism) even though the predictive accuracy of instruments has improved over the years with more research into the correlates of crime and the development of a theory of criminal behavior and effective rehabilitation (i.e., the RNR model).

There are also concerns that the use of risk and needs assessment will have a discriminatory effect on minorities. As discussed previously, the research on the applicability of currently used instruments for minorities is mixed. Some policymakers might be concerned that instruments might find minorities to be at a higher risk for recidivism than whites because of the use of static risk factors, such as criminal history, that might be more prevalent in minority communities because they are more at risk of coming into contact with the criminal justice system. While this is a valid concern, it should also be noted that many commonly used instruments consider a wide variety of dynamic risk factors that could allow all inmates to reduce their assessed risk level. Also, actuarial assessment is the norm, which makes the process of assessing each offender's risk level more objective. Before the use of actuarial assessment, decisions about who was to be assigned to which treatment program and who was to be released on parole were left to criminal justice professionals who made assessments based on their own sets of standards, which might have been influenced by overt or subconscious biases.

## **Should Certain Inmates Be Excluded from Earning Additional Time Credits?**

One issue policymakers might consider is whether certain inmates should be excluded from earning extra time credits for participating in rehabilitative programs and productive activities. Some legislative proposals would exclude inmates who were convicted of certain offenses, such as violent and sex offenses, from earning additional time credits for participating in rehabilitative programming.<sup>65</sup> Research suggests that inmates should be assessed for risk and decisions about programming and supervision should be made based on those assessments regardless of the inmate's current offense. However, it might be argued that inmates who are convicted of serious offenses, such as violent or sex offenses, should not be eligible to be released from prison early, regardless of what they do to reduce their risk of recidivism.

Another issue that policymakers might consider is whether excluding inmates convicted for certain offenses would have a disparate effect on racial or ethnic minorities. Some policymakers might be concerned that excluding inmates convicted of certain offenses from being eligible to receive additional time credits under the proposed assessment system might mean that inmates of color would be more likely to have to serve more time in prison. However, this would only be true to the extent that inmates of color are more likely to be convicted of offenses that would make inmates ineligible to receive additional time credits. Data available through the Bureau of

---

<sup>65</sup> See, for example, S. 467, S. 2123, H.R. 759, and H.R. 2944.

Justice Statistics' Federal Criminal Case Processing Statistics program is not detailed enough to allow CRS to analyze the potential disparate effects of the exclusions listed in the current legislative proposals. Congress might consider whether it wants to ask the U.S. Sentencing Commission or the BOP to assess the potential effects of excluding inmates convicted for certain offenses.

## **Should Priority Be Given to High-Risk Offenders?**

Policymakers might consider whether the proposed risk and needs assessment system should focus on high-risk inmates. The RNR principles state that high-risk individuals should be the focus of interventional programming.

Research on the risk principle suggests that recidivism is only reduced when high-risk offenders are placed in programs where they receive intensive levels of services.<sup>66</sup> In some instances, research also found that low-risk offenders who were placed in intensive treatment programs actually had an increased likelihood of recidivism.<sup>67</sup> This could be because placing low-risk offenders in intensive programming interrupts support structures or self-correcting behaviors that already exist, or because it exposes low-risk offenders to high-risk offenders who may have a negative influence on low-risk offenders' thoughts or behaviors.<sup>68</sup>

Some legislative proposals would require the BOP to phase-in the risk and needs assessment system.<sup>69</sup> During the phase-in period, low-risk prisoners would be given priority for programs and activities over moderate- and high-risk prisoners. In addition, higher-risk inmates would be required to participate in more rehabilitative programming, but inmates with low or no risk of recidivating would also be required to participate in rehabilitative programming. Other legislative proposals would require inmates who are deemed to be low risk and without need of recidivism reduction programming to continue to participate in productive activities.<sup>70</sup> Policymakers might consider whether inmates who are deemed to be low risk should immediately be placed in prerelease custody in order to open spots for moderate- and high-risk inmates who are in need of rehabilitative programming.

## **Should Risk and Needs Assessment Be Used in Sentencing?**

Another issue policymakers might consider is whether risk and needs assessment should be used in sentencing to help identify low-risk offenders who could be diverted to community supervision rather than incarcerated. As discussed previously, research suggests that low-risk offenders should not be subjected to intensive treatment (and some research indicates that it might be criminogenic) and they might be able to be effectively supervised in the community. Some legislation would require the BOP, to the extent practicable, to house low-risk inmates together, which might help reduce the criminogenic effects of placing low-risk offenders in prison.<sup>71</sup> Legislative proposals would also seek ways to try to place some inmates in prerelease custody earlier.<sup>72</sup> However, if the purpose of the legislation is to reduce the federal prison population and

---

<sup>66</sup> *The Psychology of Criminal Conduct*, p. 48.

<sup>67</sup> *Ibid.*

<sup>68</sup> Vera Institute of Justice's memorandum re: risk and needs assessment, p. 2.

<sup>69</sup> See, for example, H.R. 759.

<sup>70</sup> See, for example, S. 467 and S. 2123.

<sup>71</sup> See, for example, S. 467 and S. 2123.

<sup>72</sup> See, for example, S. 467, S. 2123, H.R. 759, and H.R. 2944.

save money, it is significantly cheaper to place offenders on probation compared to incarcerating them.<sup>73</sup>

While some scholars have argued for integrating risk assessment into sentencing guidelines to help judges determine the appropriate sentences for offenders,<sup>74</sup> research suggests that if such assessment were to be integrated into sentencing, it might be best to use it as a way to screen-out low-risk offenders. Three researchers who conducted a meta-analysis of the research on risk assessment instruments concluded that instruments could be used to make informed decisions about treatment or management of offenders.<sup>75</sup> However, the high number of false positives limits their effectiveness as a tool to make decisions about who should be sent to prison for longer periods of incarceration because they pose the greatest threat of reoffending. Simply stated, if assessment were to be used to make decisions about who should be incarcerated for long periods of time because certain offenders were at a high risk for committing more offenses, there is the potential to incarcerate a significant number of people who would not commit any more offenses. The researchers concluded that the results of their analysis “suggest that these tools can effectively screen out individuals at low risk of future offending.”<sup>76</sup>

However, the idea of using risk and needs assessment in sentencing is not without controversy. DOJ, while acknowledging the important role the use of evidence-based practices plays in effective rehabilitation programs and reentry practices, has raised concerns about making risk assessment a part of determining sentences for federal offenders.<sup>77</sup> DOJ echoes previously mentioned concerns that risk assessment bases decisions on group dynamics and that determining someone’s risk of reoffending on static risk factors might place certain groups of offenders at a disadvantage. DOJ also argues that using risk assessment in determining sentences would erode the certainty in sentencing, something Congress attempted to address when it passed the Sentencing Reform Act (P.L. 98-473), which eliminated parole for federal inmates and established a determinate sentencing structure under the federal sentencing guidelines. Certainty in sentencing, argues DOJ, is a key factor in deterring crime. DOJ also argues that sentencing should primarily be about holding offenders accountable for past criminal behavior.

## Should There Be a Decreased Emphasis on Punishment?

If Congress were to consider legislation to implement risk and needs assessment in the federal prison system, policymakers might consider whether implementing a policy of making decisions based on an offender’s risk level is compatible with a perceived desire to continue to incarcerate certain offenders for as long as possible. Some legislation would exempt inmates convicted of certain crimes from being eligible from earning extra time credits.<sup>78</sup> This would mean that

---

<sup>73</sup> The Administrative Office of the U.S. Courts reports that in 2012 the average annual cost of probation supervision was \$3,347 per probationer, compared to \$28,948 to house an inmate in a federal prison. Administrative Office of the U.S. Courts, “Supervision Costs Significantly Less than Incarceration in Federal System,” July 18, 2013, <http://news.uscourts.gov/supervision-costs-significantly-less-incarceration-federal-system>.

<sup>74</sup> Jordan M. Hyatt, Mark H. Bergstrom, and Steven L. Chanenson, “Follow the Evidence: Integrate Risk Assessment into Sentencing,” *Federal Sentencing Reporter*, vol. 23, no. 4, April 2011, pp. 266-268.

<sup>75</sup> Seena Fazel, Jay P. Singh, and Helen Doll, “Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24,827 People: Systematic Review and Meta-analysis,” *BMJ: British Medical Journal*, vol. 345, July 24, 2012, <http://www.bmj.com/content/bmj/345/bmj.e4692.full.pdf>.

<sup>76</sup> *Ibid.*, p. 4.

<sup>77</sup> Letter from Jonathan J. Wroblewski, Director, Office of Policy and Legislation, Criminal Division, U.S. Department of Justice, to The Honorable Patti B. Saris, Chair, U.S. Sentencing Commission, July 29, 2014.

<sup>78</sup> See, for example, S. 467, S. 2123, H.R. 759, and H.R. 2944.

offenders convicted of certain offenses would be required to serve a greater proportion of their sentences in prison even if they are deemed to be at a low risk for recidivism. As discussed previously, it is an offender's past history of antisocial behavior, and not the offender's current offense, that is indicative of a risk for recidivism. Therefore, the policy of requiring certain offenders to serve most of their sentences in prison might, in some capacity, undermine the potential effectiveness of a risk and needs assessment system.

Research has questioned the effectiveness of incarceration as a way to reduce crime. It suggests that while incarceration did contribute to lower violent crime rates in the 1990s, there are declining marginal returns associated with ever-increasing levels of incarceration.<sup>79</sup> The diminishing level of return resulting from higher levels of incarceration might be explained by the fact that higher levels of incarceration are likely to include more offenders who are either at the end of their criminal careers or who were at a low risk of committing crimes at a high rate (so-called "career criminals").<sup>80</sup> Another possible reason for diminishing marginal returns might be that more of the individuals incarcerated over the past three decades have been incarcerated for crimes where there is a high level of replacement (i.e., incarcerating one offender "opens the market" for a new offender to take that person's place).<sup>81</sup> For example, if a drug dealer is incarcerated and there is no decrease in demand for drugs in the drug market, it is possible that someone will step in to take that person's role; therefore, no further crimes may be averted by incarcerating the individual. It is also possible that being imprisoned with other offenders is actually criminogenic, especially for low-risk offenders.<sup>82</sup>

Research on the psychology of punishment also provides insight into why incarceration might provide a limited deterrent effect. For punishment to be successful at suppressing behavior it requires

- the immediate delivery of an intense level of punishment,
- catching and punishing criminals for every offense,
- not allowing the offender to be able to escape from the consequences of the behavior,
- making the density of the punishment associated with the behavior greater than the density of the rewards, and
- the punishment be consistent with the characteristics of the offender.<sup>83</sup>

However, "the necessary conditions for effective punishment are virtually impossible to meet for the criminal justice system. Police cannot be everywhere to ensure the certainty of detection, the courts cannot pass sentence quickly enough, and correctional officials have difficulties ensuring adequate supervision and monitoring."<sup>84</sup>

---

<sup>79</sup> Anne Morrison Piehl and Bert Useem, "Prisons," in *Crime and Public Policy*, ed. Joan Petersilia and James Q. Wilson, 2<sup>nd</sup> ed. (New York: Oxford University Press, 2011), p. 542.

<sup>80</sup> Doris Layton MacKenzie, "Reducing the Criminal Activities of Known Offenders and Delinquents: Crime Prevention in the Courts and Corrections," in *Evidence-based Crime Prevention*, ed. Lawrence W. Sherman, David P. Farrington, Brandon C. Welsh, and Doris Layton MacKenzie (New York: Routledge, 2002), p. 337.

<sup>81</sup> Bert Useem and Anne Morrison Piehl, *Prison State: The Challenge of Mass Incarceration* (New York: Cambridge University Press, 2008), p. 74.

<sup>82</sup> *The Psychology of Criminal Conduct*, p. 433.

<sup>83</sup> *Ibid.*, pp. 443-447.

<sup>84</sup> *Ibid.*, p. 451.

There is also an argument to be made about the purpose of incarceration. While there might be a minimal general deterrent effect associated with incarceration, it does provide for incapacitation, which can reduce the number of crimes an incarcerated offender can commit. Also, long prison terms might provide for society's sense of justice. Sentencing someone to prison for several years, or even decades, could be viewed as a way for society to say that there are certain behaviors that will not be tolerated, and those who commit such transgressions deserve to receive severe punishment for them.

## Appendix A. Comparison of Risk and Needs Assessment Legislation

This appendix provides a comparison of the risk and needs assessment-related provisions in four bills introduced in the 114<sup>th</sup> Congress:

- S. 467, the CORRECTIONS Act;
- S. 2123, the Sentencing Reform and Corrections Act of 2015;
- H.R. 759, the Recidivism Risk Reduction Act; and
- H.R. 2944, the Sensenbrenner-Scott SAFE Justice Reinvestment Act of 2015.

The text of S. 2123 generally incorporates the text of S. 467, with a few key differences, highlighted below.

### Establishment of an Assessment System

S. 467 and S. 2123 would require the Department of Justice (DOJ) to establish, within 30 months of the enactment of the bill, a Post-Sentencing Risk and Needs Assessment System (Assessment System) for use in the BOP that would

- assess and determine the recidivism risk level of all inmates and classify each inmate as being at low, moderate, or high risk for recidivism;
- to the extent practicable, determine the risk of violence for all inmates;
- ensure that, to the extent practicable, low-risk inmates are housed and assigned to programs together;
- assign inmates to rehabilitative programs and productive activities based on their risk level and criminogenic needs;
- periodically reassess and update an inmate’s risk level and programmatic needs; and
- provide information on best practices concerning the tailoring of rehabilitative programs to the criminogenic needs of each inmate.

H.R. 759 would also require DOJ to develop and release an Assessment System for use by the BOP, but it would require DOJ to establish the system within 180 days of the bill becoming law. The requirements for the Assessment System under H.R. 759 are similar to those of S. 467, but H.R. 759 would *not* require the Assessment System to determine the risk of violence for all inmates, nor require that low-risk inmates be housed together and assigned to the same programs.

H.R. 2944 would require DOJ to develop an Assessment System within one year of the bill becoming law. The requirements for the system that would be established under H.R. 2944 are similar to those of the other two bills in that H.R. 2944 would require the system to be used to assess and determine the risk and needs factors for federal inmates and to assign inmates to recidivism reduction programs based on their risk and needs. The Assessment System that would be established by the bill would not be required to assess each inmate’s risk of violence nor require low-risk inmates to be segregated. However, the bill notes that “some activities or excessive programming may be counter-productive for some prisoners” and as such, it would allow DOJ to provide guidance to the BOP on the quality and quantity of rehabilitative programming that is both appropriate and effective.



All four pieces of legislation would require DOJ, when developing the Assessment System, to use the best available research and best practices in the field of risk and needs assessment. S. 467, S. 2123, and H.R. 759 would allow DOJ to develop its own instrument or use an existing instrument. H.R. 2944 would require DOJ to prescribe a “suitable intake assessment tool” but it is silent as to whether the instrument would need to be developed in-house or if an existing instrument could be used. In addition, all four bills would require DOJ either to validate the instrument on the federal prison population or to ensure that the instrument has been validated using federal inmates.

S. 2123 would also require DOJ to make adjustments to the system on a regular basis, but not less than once every three years. In doing so, DOJ would be required to consider the best evidence available on effective means of reducing recidivism rates and to make adjustments, to the extent possible, to ensure that the system does not result in any unwarranted disparities, including disparities amongst similarly classified inmates of different racial groups. S. 2123 would require DOJ to adjust the system to reduce disparities to the greatest extent possible. The bill would also require DOJ to coordinate with the U.S. Probation and Pretrial Services Office to ensure that the findings of each offender’s presentence report are available and considered in the Assessment System.

## **Expanding Rehabilitative Programs**

S. 467 and S. 2123 would require the BOP, subject to the availability of appropriations, to make recidivism reduction programs and productive activities available to all eligible inmates within six years of enactment of the legislation. Both bills would also require the National Institute of Corrections to evaluate all programs and activities to ensure that they are evidence based and effective at reducing recidivism.

H.R. 2944 would require the BOP, subject to the availability of appropriations, to make recidivism reduction programs and productive activities available to all eligible inmates within one year of enactment.

H.R. 759 would also require the BOP to expand, subject to appropriations, recidivism reduction programs and productive activities for inmates. However, H.R. 759 would phase in expansion of programs and activities. The BOP would be required to provide rehabilitative programming and productive programs to 20% of inmates within one year of the date when risk and needs assessments are completed for all inmates. The BOP would be required to provide rehabilitative programming and productive activities to an additional 20% of inmates each year until they are serving all inmates. During the phase-in period, low-risk inmates would be given first priority for participation in rehabilitative programs and productive activities. Moderate- and high-risk inmates would be given second and third priority, respectively. Also, within risk levels, priority would be given to inmates who are closer to finishing their sentences.

All four bills would allow the BOP to enter into partnerships with nonprofit organizations, educational institutions, and private entities in order to provide rehabilitative programs and activities for inmates. S. 2123 would also allow the BOP to enter into partnerships with “industry-sponsored organizations that deliver workforce development and training that lead to recognized certification and employment.”

## **Assessing the Risk and Needs of Inmates**

S. 467 and S. 2123 would require the BOP to conduct an initial risk and needs assessment for all inmates within 30 months of the bill becoming law. Both bills would also require the BOP to

reassess each inmate at least once a year for inmates within three years of release; at least once every other year for inmates who are within 10 years of release; and at least once every three years for every other inmate.

H.R. 759 would require the BOP to periodically reassess inmates who successfully participate in rehabilitative programs and productive activities (with high- and moderate-risk inmates receiving more frequent evaluations) and assign inmates to the proper programs and activities if their risk levels change.

H.R. 2944 would require the BOP to develop a case plan for each inmate that targets each inmate's risk and needs and helps guide the inmate's rehabilitation. Case plans would have to be completed within 30 days of an inmate's initial admission. Case plans would be required to

- include programming and treatment requirements based on the inmate's assessed risk and needs;
- ensure that inmates whose risk and needs do not warrant recidivism reduction programming participate in and successfully complete productive activities, including prison jobs; and
- ensure that eligible inmates participate in and successfully complete recidivism reduction programming or productive activities throughout their entire term of incarceration.

H.R. 2944 would require the BOP to provide each inmate with a copy of the case plan and discuss the case plan with the inmate. The BOP would be required to review the case plan with the inmate every six months to assess the inmate's progress towards completing it and whether the inmate needs to participate in additional or different rehabilitative programs.

## **Training for Staff on Using the Assessment System**

All four bills would require BOP staff who are responsible for administering the Assessment System to be trained on how to properly use the system, which includes a requirement that staff demonstrate competence in administering the instrument. S. 467, S. 2123, and H.R. 759 would require DOJ to monitor and assess the use of the Assessment System and to periodically audit the use of the system in BOP facilities. H.R. 2944 would require DOJ, the Government Accountability Office, and DOJ's Inspector General's Office to monitor and assess the use of the Assessment System and to conduct separate and independent periodic audits of the use of the system.

## **Additional Time Credits and Other Incentives**

S. 467 and S. 2123 would grant additional time credit for inmates who successfully complete 30 days of rehabilitative programming and productive activities. Every inmate would be eligible to earn five additional days of credit upon completion. Inmates who are deemed low risk would be eligible to receive an additional five days. S. 467 would exempt the following inmates from earning additional time credits:

- inmates serving a sentence for a second federal offense;
- inmates who were in the highest criminal history category under the U.S. Sentencing Guidelines at the time of sentencing; and

- any inmate sentenced for a terrorism offense,<sup>85</sup> a crime of violence,<sup>86</sup> a sex offense,<sup>87</sup> racketeering,<sup>88</sup> engaging in a continuing corrupt criminal enterprise,<sup>89</sup> a federal fraud offense for which the inmate was sentenced to more than 15 years' imprisonment, or a crime involving child exploitation.<sup>90</sup>

S. 2123 would exempt the following inmates from earning additional time credits:

- inmates serving a sentence for a second federal offense, which would not include any offense under the Major Crimes Act (relating to federal jurisdiction over certain enumerated crimes committed by Native Americans on tribal lands) for which the offender was sentenced to less than 13 months;
- inmates who have 13 or more criminal history points, as determined under the U.S. Sentencing Guidelines, at the time of sentencing, unless the court determines in writing that the defendant's criminal history score substantially over-represents the seriousness of the offender's criminal history or the likelihood that the offender will commit other crimes;
- any inmate sentenced for a terrorism offense,<sup>91</sup> a crime of violence,<sup>92</sup> a sex offense,<sup>93</sup> engaging in a continuing corrupt criminal enterprise,<sup>94</sup> or a federal fraud offense for which the inmate was sentenced to more than 15 years' imprisonment, a crime involving child exploitation;<sup>95</sup> or
- inmates convicted of offenses under chapter 11 (relating to bribery, graft, and conflicts of interest); chapter 29 (relating to elections and political activities); chapter 63 (involving a scheme or artifice to deprive someone of the intangible right of honest services); chapter 73 (relating to the obstruction of justice); chapter 95 or 96 (relating to racketeering and racketeering influenced and corrupt organizations); chapter 110 (relating sexual abuse and other abuse of children); or sections 1028A, 1031, or 1040 (relating to fraud) of Title 18 of the U.S. Code.

H.R. 759 would also allow inmates to earn additional time credits for successfully participating in rehabilitative programs or productive activities, but the credit structure would be different. Under H.R. 759, low-risk inmates would be eligible to receive 30 days of time credits for each month they successfully participate in a rehabilitative program or productive activity; moderate-risk inmates would be eligible to receive 15 days, and high-risk inmates would be eligible to receive 8 days. H.R. 759 lists 47 offenses that would make federal inmates ineligible to receive additional time credits for participating in rehabilitative programs or productive activities. The enumerated offenses could generally be classified as violent offenses, terrorism offenses, espionage offenses,

---

<sup>85</sup> As defined at 18 U.S.C. §2332b(g)(5).

<sup>86</sup> As defined at 18 U.S.C. §16.

<sup>87</sup> As described in 42 U.S.C. §16911.

<sup>88</sup> As defined at 18 U.S.C. §1962.

<sup>89</sup> As defined at 21 U.S.C. §848.

<sup>90</sup> As defined at 42 U.S.C. §17601.

<sup>91</sup> As defined at 18 U.S.C. §2332b(g)(5).

<sup>92</sup> As defined at 18 U.S.C. §16.

<sup>93</sup> As described in 42 U.S.C. §16911.

<sup>94</sup> As defined at 21 U.S.C. §848.

<sup>95</sup> As defined at 42 U.S.C. §17601.

human trafficking offenses, sex and sexual exploitation offenses, and high-level drug offenses.<sup>96</sup> The bill would also exclude inmates with three or more convictions for crimes of violence or drug trafficking offenses.

H.R. 2944 would allow inmates to earn 10 days of time credits for each month they successfully comply with their case plans. Unlike the other two pieces of legislation, under H.R. 2944 all inmates would be eligible to receive the same amount of time credits, regardless of risk score. Also, unlike the two other pieces of legislation, H.R. 2944 would allow the BOP to retroactively award time credits to eligible inmates for participating in rehabilitative programs and activities before enactment of the bill. Inmates who have been convicted of murder,<sup>97</sup> terrorism,<sup>98</sup> or sex offenses<sup>99</sup> would not be eligible to receive time credits for participating in rehabilitative programming.

S. 467, S. 2123, and H.R. 2944 would require the BOP to develop other incentives, such as additional telephone or visitation privileges, for inmates who are exempt from earning additional time credits. H.R. 759 would allow any prisoner who successfully participates in a rehabilitative program or productive activity to receive, for use with family, close friends, mentors, and religious leaders, up to 30 minutes per day and up to 900 minutes per month in phone privileges and, as determined by the facility's warden, additional visitation time.

H.R. 2944 would require the BOP to amend its inmate disciplinary program to provide for the reduction of earned time credits for inmates who violate institutional rules or the rules of the rehabilitative program or productive activity.<sup>100</sup> The amendments would be required to specify the level of violations and the corresponding penalties; that any loss of earned time credits does not apply to earning credits in the future; and a procedure for inmates to have lost time credits restored based on their progress. H.R. 759 includes a similar requirement. S. 467 and S. 2123 would allow the BOP to reduce earned time credits for misbehavior, but it would not require the BOP to do so.

Under both S. 467 and S. 2123, inmates would not be allowed to accrue the proposed additional time credits if the inmate has accrued other time credits for participation in another program under another provision of law. Under both House bills, the time credits earned for participating in rehabilitative programs and productive activities would be in addition to any other rewards or incentives for which inmates might be eligible.

## **Placement in Prerelease Custody**

The extra time credit inmates could earn under S. 467, S. 2123, and H.R. 759 would allow them to be placed on prerelease custody earlier. Under both S. 467 and S. 2123, inmates who are deemed to be at a low risk for recidivism within one year of being eligible to be placed in

---

<sup>96</sup> “High-level drug offenses” means offenses under section 401(a) of the Controlled Substances Act (21 U.S.C. 841(a)), relating to manufacturing or distributing a controlled substance, but only in the case of a conviction for an offense described in subparagraphs (A), (B), or (C) of subsection (b) of that section for which death or serious bodily injury resulted from the use of such substance.

<sup>97</sup> Only in cases where it was shown beyond a reasonable doubt that the inmate had the intent to cause death and death resulted.

<sup>98</sup> As defined at 18 U.S.C. §2332b(g)(5).

<sup>99</sup> As described in section 111 of the Sex Offender Registration and Notification Act (Title I of the Adam Walsh Child Protection and Safety Act of 2006 (P.L. 109-248)).

<sup>100</sup> For more information on BOP's inmate disciplinary program see CRS Report R42486, *The Bureau of Prisons (BOP): Operations and Budget*, by Nathan James.

prerelease custody, or inmates who are deemed at a moderate risk for recidivism but their most recent risk and needs assessment shows that their risk of recidivism has decreased, would be eligible to be placed in a residential reentry center (RRC, i.e., a halfway house) or home confinement. Inmates who are deemed to be low risk for recidivism can be placed on community supervision. Inmates who have earned less than 36 months of additional good time credit would only be eligible to spend one-half of that time on community supervision, while inmates who have earned 36 months or more of additional good time credit would be eligible to serve the amount of such credit exceeding 18 months on community supervision.

H.R. 759 would allow the BOP to place inmates who are deemed to be low risk, who have earned time credits equal to the amount of time remaining on their sentences, and who are otherwise deemed qualified, in prerelease custody. All inmates transferred to prerelease custody would be placed on home confinement. Inmates would be required to remain on home confinement until they served at least 85% of their imposed sentence.

Under both Senate bills, any period of supervised release imposed on an inmate would be reduced by the amount of time the prisoner spent in prerelease custody. Inmates would not be eligible to be transferred to community supervision unless the amount of time the inmate could spend on community supervision is equal to or greater than the amount of time remaining on the inmate's period of prerelease custody.

H.R. 2944 does not contain any provisions related to special conditions for inmates placed on prerelease custody pending completion of their sentences.

## **Judicial Review of Prerelease Custody Placement**

Both S. 467 and S. 2123 would not allow the BOP to transfer any inmate sentenced to more than three years of incarceration to prerelease custody unless the BOP provides notice to the U.S. Attorney's office in the district where the inmate was convicted. The federal government would be allowed to challenge an inmate's prerelease custody. A court would be allowed to deny an inmate's transfer to prerelease custody or modify the terms of such transfer if, after conducting a hearing, the court finds by a preponderance of the evidence that placing the inmate on prerelease custody is inconsistent with the factors specified in paragraphs (2), (6), and (7) of 18 U.S.C. §3553(a).

H.R. 759 would require the BOP to notify the court in the district in which the inmate was convicted of its intention to place the inmate in prerelease custody. A judge would be required to approve or deny the recommendation within 30 days. However, the judge would only be able to deny the recommendation if he or she finds through clear and convincing evidence that the inmate's actions after conviction warrant denial of the transfer to prerelease custody. Failure of the judge to approve or deny the recommendation within 30 days would be treated as an approval.

None of the bills contain language that would allow inmates to appeal a court's decision to deny them placement in prerelease custody.

## **Appendix B. Commonly Used Risk and Needs Assessment Instruments**

There are many different risk and/or needs assessment instruments currently available. Some are only comprised of static risk factors while some use a combination of static and dynamic risk factors. Some are used to predict general recidivism while others focus on predicting recidivism for certain populations of offenders, such as sex offenders or domestic abusers. **Table B-1** presents a summary of the key aspects of seven commonly used risk and needs assessment instruments. The information provided in **Table B-1** is meant to provide examples of the differences in how some risk and needs assessment instruments are developed, the requirements to administer them, and the items they use to assess risk and needs.

**Table B-I. Commonly Used Risk and Needs Assessment Instruments**

Instrument	Background Information	Administration Requirements	Instrument Contents
<p>Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)</p>	<p>The original COMPAS system was created in the late 1990s. The instrument was designed to assess key risk and needs factors in adult and youth correctional populations and to provide decision support for practitioners charged with case planning and management. COMPAS can assess four types of risk (general recidivism, violent recidivism, non-compliance, and failure to appear). Originally developed and validated using offenders in New York, COMPAS has since been modified as revalidation data offers new insights on the performance and validity of the instrument.</p>	<p>COMPAS allows for some degree of flexibility in the administration process. Offender data collection options include offender self-report, scripted interviews, and structured interviews as part of a web-based, automated assessment process. The developer offers training that covers practical use, interpretation of results, and case planning strategies. Advanced training options are available on the theoretical underpinnings of offender assessments, gender responsivity training, motivational interviewing, and other topics.</p>	<p>The COMPAS Core assessment for adult offenders contains both static and dynamic factors. Content may be individually tailored based on jurisdictional needs and resources, but can include four risk and four need scales:</p> <ul style="list-style-type: none"> <li>• Risk: failure to appear, non-compliance (technical violations), general recidivism, violent recidivism.</li> <li>• Criminogenic needs: cognitive-behavioral, criminal associates/peers, criminal involvement, criminal opportunity, criminal personality, criminal thinking (self-report), current violence, family criminality, financial problems, history of non-compliance, history of violence, leisure/boredom, residential instability, social adjustment, social environment, social isolation, socialization failure, substance abuse, vocation/education</li> </ul>
<p>Inventory of Offender Risk, Needs, and Strengths (IORNS)</p>	<p>IORNS was created in 2006 as an offender assessment of static risk, dynamic risk/need, and protective strength factors. The tool is complemented by several subscales for specific assessments in the areas of violent and sexual criminal behavior.</p>	<p>Administrators must hold a degree in forensic or clinical psychology or psychiatry plus satisfactory completion of appropriate coursework in psychological testing, or have a license or certification from an agency that requires such training and experience. Line staff can administer the self-report assessment to offenders and score the results, but they must be supervised by a licensed professional who is also responsible for</p>	<p>IORNS is a 130-item true/false self-report questionnaire that assesses static risk, dynamic risk/need, and protective strength factors in separate indices. It consists of four total indices and eight scales.</p> <ul style="list-style-type: none"> <li>• The Static Risk Index (SRI) contains 12 criminal history items.</li> <li>• The Dynamic Need Index (DNI) contains 79 items in the form of six</li> </ul>

Instrument	Background Information	Administration Requirements	Instrument Contents
		interpreting the instrument.	<p>dynamic need scales: Criminal Orientation, Psychopathy, Intra/Interpersonal Problems, Alcohol/Drug Problems, Aggression, and Negative Social Influences.</p> <ul style="list-style-type: none"> <li>The Protective Strength Index (PSI) contains 26 items in the form of two scales: Personal Resources and Environmental Resources.</li> </ul>
<p>Level of Service Inventory-Revised (LSI-R), Level of Service/Case Management Inventory (LS/CMI), and Level of Service/Risk, Need, Responsivity (LS/RNR)</p>	<p>LSI-R was developed in 1995 and validated using a Canadian criminal population. It is a “third generation” risk and needs assessment instrument. LS/CMI is the “fourth generation” revision of LSI-R that assesses offender risk, needs, and responsivity (RNR) to inform case planning via a built-in case management system. The LS/RNR is similarly comprised of the updated risk, need, and responsivity scales, but offer these separately from the LS/CMI case management system for organizations already equipped with established case management systems of their own.</p>	<p>LSI-R and LS/CMI are administered through a structured interview between the interviewer and offender, with the recommendation that supporting documentation be collected from family members, employers, case files, drug tests, and other relevant sources as needed. Those who administer the exam must have an understanding of the principles of tests and measurements or be supervised by someone who does; a professional with advanced training in psychological assessment or a related discipline must assume responsibility for the instrument’s use, interpretation, and communication of results.</p>	<p>LSI-R and LS/CMI contain a mix of static and dynamic factors, developed from recidivism literature, professional opinions of probation officers, and relevant social learning theory on criminal behavior.</p> <p>LSI-R is a 54-item risk and needs assessment instrument that consists of 10 areas: Criminal History, Education and Employment, Financial, Family and Marital, Accommodations, Leisure and Recreation, Companions, Alcohol/Drug Problems, Emotional/Personal, and Attitudes/Orientation.</p> <p>LS/CMI refined and combined content of the LSI-R into 43 items in 8 sections: Criminal History, Education/Employment, Family/Marital, Leisure/Recreation, Companions, Alcohol/Drug Problems, Procriminal Attitude/Orientation, and Antisocial Pattern.</p> <p>LS/CMI system contains seven additional sections. Sections 2-5 of LS/CMI identify additional risk factors (personal problems; social, health, and responsivity considerations; perpetration history; mental health; procriminal</p>



Instrument	Background Information	Administration Requirements	Instrument Contents
Ohio Risk Assessment System (ORAS)	ORAS was developed in 2006 as a collaborative effort between the Ohio Department of Rehabilitation & Correction (DRC) and the University of Cincinnati Center for Criminal Justice Research (CCJR). The goal was to create a consistent, reliable, standardized system of tools that could be used at various decision points in the criminal justice system to facilitate communication and continuity across criminal justice agencies. ORAS is a “fourth generation” assessment instrument.	No specialized education is necessary to administer ORAS. However, researchers at CCJR have assembled a mandatory training package for those interested in using ORAS. ORAS uses a combination of structured interviews, official records, and other collateral sources to complete the assessment instrument. Offenders also complete a self-report questionnaire to supplement this information.	<p>attitude/orientation; incarceration history, and concerns). Sections 6-7 provide a summary of risks and needs, allowing for clinical overrides of assessment recommendations based on atypical offender situations. Section 8 provides tools for program and placement decisions.</p> <p>ORAS consists of 101 items divided between six tools. All tools contain both static and dynamic factors. The tools in ORAS are</p> <ul style="list-style-type: none"> <li>• Pretrial Assessment Tool;</li> <li>• Community Supervision Screening Tool;</li> <li>• Community Supervision Tool: assesses criminal history, education, employment, and financial situation, family and social support, neighborhood problems, substance use, peer associations, and criminal attitudes and behavioral patterns;</li> <li>• Prison Screening Tool;</li> <li>• Prison Intake Tool (PIT): assesses age, criminal history, school behavior and employment, family and social support, substance abuse and mental health, and criminal lifestyle; and</li> <li>• Prison Reentry Tool: assesses age, criminal history, social bonds, and criminal attitudes and behavioral patterns.</li> </ul>

Instrument	Background Information	Administration Requirements	Instrument Contents
Offender Screening Tool (OST)	<p>In 1998, the Maricopa County (Arizona) Adult Probation Department (MCAPD), working with consultant Dr. David Simourd, developed and implemented its own assessment instrument, the Offender Screening Tool (OST). MCAPD originally sought to create a risk/needs tool that would (1) provide a broad, overall assessment of offender risk/needs, (2) incorporate static and dynamic risk factors most predictive of criminal behavior, (3) provide information that could be used to determine risk of recidivism and guide case planning/management decisions, and (4) be meaningful and valuable to staff. As a greater variety of cognitive-behavioral treatment programs became available in the county, Dr. Simourd and MCAPD expanded OST to include additional needs domains. OST was implemented statewide in 2005.</p>	<p>OST is administered at the presentencing stage by interviewers who enter information into a computerized system for automated scoring. No specialized certifications are required, but all staff members receive training. In Maricopa County, the presentence division receives training on how to administer and interpret results from OST; all other probation department staff receive training on interpretation and how to use results to inform case planning and management.</p>	<p>The OST contains 44 items (14 static, 30 dynamic) in 10 domains:</p> <ul style="list-style-type: none"> <li>• Vocational/Financial,</li> <li>• Education,</li> <li>• Family and Social Relationships,</li> <li>• Residence and Neighborhood,</li> <li>• Alcohol,</li> <li>• Drug Abuse,</li> <li>• Mental Health,</li> <li>• Attitude, and</li> <li>• Criminal Behavior.</li> </ul> <p>The final domain, Physical Health/Medical, is used exclusively as a responsivity factor.</p>
Static Risk and Offender Needs Guide (STRONG)	<p>In 1999, the Washington Legislature directed the Department of Corrections (DOC) to improve the classification of felony offenders and to deploy staff and rehabilitative resources more effectively. The Washington State Institute for Public Policy (WSIPP) examined the validity of the risk instrument the DOC was using at the time (LSI-R) and thought that the predictive power of the assessment could be improved by including more static risk items. WSIPP, at the behest of DOC, created a new static risk instrument (Static Risk Assessment) comprised of only offender demographic and criminal history</p>	<p>The Static Risk Assessment is conducted based on a thorough investigation of offender criminal history information. No offender interview is necessary. No specialized administrator qualifications are required to administer the Offender Needs Assessment; staff members may conduct the structured interview. It is recommended that line staff complete routine booster training sessions in addition to an initial training program for quality assurance purposes. For improved quality control, Washington established a small, dedicated intake unit to conduct all risk assessments statewide.</p>	<p>STRONG consists of two separate assessments. The Static Risk Assessment is conducted first based on the offender's criminal history information and contains 26 items in the following domains: demographics, juvenile record, commitment to the DOC, total adult felony record, total adult misdemeanor record, and total sentence/supervision violations.</p> <p>Calculated separately, the Offender Needs Assessment contains 55 items in 10 domains: education, community employment, friends, residential, family, alcohol/drug use, mental health, aggression, attitudes/behaviors, and</p>

Instrument	Background Information	Administration Requirements	Instrument Contents
	<p>information, which was completed in 2006. In 2008, DOC implemented their automated offender assessment and case planning system. This automated system included the Static Risk Assessment and an Offender Needs Assessment, which is used to identify offender needs and protective factors for use in case planning. STRONG is considered a “fourth generation” risk and needs assessment instrument.</p>		<p>coping skills.</p>
<p>Wisconsin Risk/Needs Scales (WRN) and Correctional Assessment and Intervention System (CAIS)</p>	<p>The Wisconsin Classification System was created in 1977. This system is comprised of the Wisconsin Risk/Needs scales (WRN) and the Client Management Classification (CMC) responsivity and case management tool. To facilitate practitioner use of the system, the National Council on Crime and Delinquency (NCCD) updated the tools in 2004 and created the automated, web-based Correctional Assessment and Intervention System (CAIS).</p>	<p>No specialized education is required; trained line staff can administer WRN or CAIS. NCCD developed and administers a training package for the CAIS tool.</p>	<p>WRN is a 53-item interview-driven assessment. Content areas include criminal history, education/employment, family/friends, mental/emotional stability, plans/problems, health, sexual behavior, drug/alcohol usage, and financial management. The CMC is a 71-item interview-based case planning process that categorizes offenders into one of four possible typologies (Selective Intervention, Casework/Control, Environmental Structure, and Limit Setting). These classifications can then be used to guide case planning strategies.</p> <p>CAIS is an automated assessment and case management system that includes an updated version of WRN and CMC. A new risk and needs tool was created based on the results of a meta-analysis and can be included in CAIS.</p>

**Source:** CRS presentation of information provided in Appendix A to Pamela M Casey, Roger K. Warren, and Jennifer K. Elek, *Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group*, National Center for State Courts, Williamsburg, VA, 2011.

## **Author Contact Information**

Nathan James  
Analyst in Crime Policy  
njames@crs.loc.gov, 7-0264

# **Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States**

**Sarah L. Desmarais, Ph.D.**

*Department of Psychology, North Carolina State University*

**Jay P. Singh, Ph.D.**

*Department of Justice, Psychiatric/Psychological Service, Canton of Zürich, Switzerland*

March 27, 2013

## **ACKNOWLEDGMENTS**

We gratefully acknowledge the research assistance and contributions of Kiersten Johnson, Krystina Dillard and Rhonda Morelock to this report, as well as Grace Seamon for her research assistance. We also thank Mr. David D'Amora and Dr. Fred Osher for their guidance in the preparation of this report.

This project was funded by the Council of State Governments Justice Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsor.

**TABLE OF CONTENTS**

---

<b>Content</b>	<b>Page</b>
Acknowledgments	i
Table of Contents	ii
List of Tables	iii
Executive Summary	1
Background	3
Issues in Risk Assessment	4
Methods of the Current Review	9
Summary of Findings across Instruments	14
Summary of Findings by Instrument	28
Other Types of Instruments Used to Assess Recidivism Risk	44
Conclusion	49
Bibliography	53
Appendix A: List of Jurisdiction-Specific Risk Assessment Instruments	57
Appendix B: Glossary of Terms	59

---

**LIST OF TABLES**

---

<b>Table</b>	<b>Page</b>
1. Criteria Used to Determine Practical Significance of Aggregate Inter-Rater Reliability Findings	12
2. Criteria Used to Determine Practical Significance of Aggregate Predictive Validity Findings	12
3. Characteristics of Risk Assessment Instruments	14
4. Types of Items Included in the Risk Assessments Instruments	15
5. Content Domains Assessed across the Risk Assessment Instruments	16
6. Characteristics of the Assessment Process Used in Studies Included in this Review	18
7. Design Characteristics and Procedures of Studies Included in this Review	20
8. Summary of Predictive Validity Findings by Performance Indicator across Studies	22
9. Validity of Total Scores in Predictive Different Forms of Recidivism	24
10. Validity of Risk Classifications in Predicting Different Forms of Recidivism	25
11. Validity of Total Scores in Predicting Recidivism by Offender Sex	26

---



## EXECUTIVE SUMMARY

### *Overview*

The rates of crime, incarceration and correctional supervision are disproportionately high in the U.S. and translate into exorbitant costs to individuals, the public and the state. Though many offenders recidivate, a considerable proportion do not. Thus, there is a need to identify those offenders at greater risk of recidivism and to allocate resources and target risk management and rehabilitation efforts accordingly. Doing so necessitates accurate and reliable assessments of recidivism risk. There is overwhelming evidence to suggest that assessments of risk completed using structured approaches produce estimates that are both more accurate and more consistent across assessors compared to subjective or unstructured approaches. More and more, structured risk assessment approaches are being used in correctional agencies.

In this review, we summarize the research conducted in the United States examining the performance of instruments designed to assess risk of recidivism, including committing a new crime and violating of conditions of supervision, among adult offenders. We focus specifically on performance of tools validated and currently used in correctional settings in the United States.

### *Methodology*

We identified instruments designed to assess risk of recidivism by searching academic research databases and Google. We identified additional instruments by looking through the reference lists of recent publications and through discussion with colleagues. Criteria for instruments to be included in the review were: a) designed to assess the likelihood of general recidivism (i.e., new offenses and violation of conditions); b) intended for assessing adult offenders (18 years of age and older); c) used in correctional settings in the United States; and d) validated in the United States. Instruments were excluded from our review if they: a) were designed to assess the likelihood of adverse outcomes for specific offenses (e.g., sexual offenses, violent offenses, spousal assault); b) were intended for assessing juvenile offenders (less than 18 years of age); c) were not used in correctional settings in the United States; d) had not been validated in the United States.; or e) were developed for use in a specific institution or ward.

We then identified studies examining the validity of these instruments using the same databases, search engine and secondary sources as above, using both the acronyms and full names of the instruments as search criteria. We searched for studies published between 1970 and 2012 in peer-reviewed journals, as well as government reports, doctoral dissertations, and Master's theses. Using this search strategy, an initial total of 173 records was filtered to a final count of 53 studies, representing 72 unique samples.

Information about the characteristics of the instruments, assessment process, and studies was collected. We also recorded information on inter-rater reliability and predictive validity, overall and by offender sex, race/ethnicity, study context, and recidivism outcome, where possible.

## ***Findings***

There were very few U.S. evaluations examining the predictive validity of assessments completed using instruments commonly used in U.S. correctional agencies. In most cases, validity had only been examined in one or two studies conducted in the United States, and frequently, those investigations were completed by the same people who developed the instrument. Also, only two of the 53 studies reported evaluations of inter-rater reliability. There was no one instrument that emerged as systematically producing more accurate assessments than the others. Performance within and between instruments varied depending on the assessment sample, circumstances, and outcome.

Some instruments performed better in predicting particular recidivism outcomes than others. Other instruments were developed to assess for specific populations (e.g., parolees) or appeared to perform better for some subgroups of offenders than others (e.g., male versus female offenders). Finally, the information and amount of time required to complete assessments varied considerably. Some instruments could be completed based solely on offender self-report; other instruments used information derived from a variety of sources, including self-report, interview, and review of official records. Still other instruments could be completed based on file review alone. The number of items included the instruments also varied considerably: from four to 130.

## ***Conclusion***

When deciding which recidivism risk assessment instrument to implement in practice, we recommend first narrowing the potential risk assessment instruments by answering the following questions: *What is your outcome of interest? What is your population? What resources are required to complete the assessment?* We then recommend careful consideration of the research evidence, including the amount and strength of the empirical support for inter-rater reliability and predictive validity, generalizability of findings, and possible sources of bias that may have impacted results. Finally, it is important to remember that the goal of risk assessment is not simply predict the likelihood of recidivism, but, ultimately, to reduce the risk of recidivism. To do so, the risk assessment tool must be implemented in a sustainable fashion with fidelity; findings of the risk assessment must be communicated accurately and completely; and, finally, information derived during the risk assessment process must be used to guide risk management and rehabilitation efforts.

## BACKGROUND

### *Prevalence of General Offending and Recidivism in the U.S.*

The crime rate in the U.S. is high, estimated at 3,295 crimes per 100,000 residents in 2011 (FBI, 2012). With 743 in 100,000 U.S. adults incarcerated at the end of 2009 (Glaze, 2011), the rate of incarceration is over four times the rate found in more than half of the world's countries (Walmsley, 2010). Indeed, though the U.S. has less than 5% of the global population, it has more than 25% of the world's prisoners (Liptak, 2008). Further, approximately one out of every 30 adults is under some form of correctional supervision (Pew Center on the States, 2009).

These high rates of crime, incarceration and correctional supervision translate into exorbitant costs. Approximately \$74 billion was spent on corrections in 2007 (Kyckelhahn, 2012). When both direct and indirect costs are considered, estimates of annual costs have reached as high as \$1.7 trillion (Anderson, 1999). Though almost two-thirds of offenders recidivate following release, another third do not go on to reoffend (Langan & Levin, 2002). Criminal justice expenditures, however, typically are distributed equally among offenders, regardless of risk level. It would be more cost-effective to allocate funding based on consideration of other factors, such as risk of recidivism and treatment needs. Indeed, correctional programs that adhere to the Risk-Need-Responsivity (RNR) model for offender assessment and rehabilitation have increased efficacy in reducing recidivism (e.g., Lowenkamp, Pealer, Smith & Latessa, 2006).

The RNR model represents an idiographic approach to risk management and rehabilitation. First, the *risk* principle dictates that treatment and intervention should be proportionate to each offender's recidivism *risk*, with more restrictive and intensive efforts used for high-risk offenders. The *need* principle calls for consideration of individual criminogenic needs to tailor treatment to each offender. Finally, the *responsivity* principle requires adapting treatment according to the individual offenders' learning styles, motivation, personalities and strengths, and use of approaches that are known to be responsive to the identified needs (Bonta & Andrews, 2007). Adherence to the principles of the RNR model necessitates accurate and reliable assessments of recidivism risk.

## ISSUES IN RISK ASSESSMENT

### *Risk Assessment in Correctional Settings in the U.S.*

Risk assessment can be defined as the process of estimating the likelihood of future offending to identify those at higher risk and in greater need of intervention. Conducting risk assessments also may assist in the identification of treatment targets and the development of risk management and treatment plans. There is overwhelming evidence to suggest that assessments of risk completed using structured approaches produce estimates that are both more accurate and more consistent across assessors compared to subjective or unstructured approaches (Ægisdóttir et al., 2006). Importantly, the use of structured approaches to classify higher risk individuals within the general offender population also produce better outcomes compared to unstructured approaches (Mamalian, 2011). More and more, correctional agencies are recommending—and many now require—the use of structured risk assessment approaches (Skeem & Monahan, 2011).

### *Evolution of Risk Assessment*

The focus and structure of risk assessment tools have shifted significantly over time. The general characteristics of four distinct generations are summarized below.

#### *First Generation*

The first generation of risk assessment is best described as unstructured professional judgment, in which the assessor relies on their professional training and information gathered from the offender, official records or other sources to inform their evaluation of risk for recidivism. It is “unstructured” insofar as there is no set checklist or protocol for completing the risk assessment, though assessors may indeed complete structured interviews during the risk assessment process. This method of assessment was widely accepted for decades prior to the development of structured risk assessment tools in the 1970s. Today, it is less frequently used, but nonetheless remains a prominent risk assessment strategy, despite evidence that accuracy of unstructured assessments risk are less accurate than chance.

#### *Second Generation*

Following decades of research focused on identifying factors that increase risk of recidivism, second generation tools represent a drastic advance in risk assessment technology. Second tools are actuarial in nature and comprised primarily of historical and static factors (e.g., sex, age and criminal history). Rather than subjective judgments of recidivism risk, instruments such as the Salient Factor Score (SFS) and Violent Risk Appraisal Guide (VRAG) instead guide assessors to consider a set list of risk factors to arrive at a numerical risk of recidivism. Actuarial instruments are described more fully in the following section.

### *Third Generation*

The third generation of risk assessment is characterized by the development of tools that include dynamic factors and criminogenic needs, and may use an actuarial or structured professional judgment approach. Third generation tools, such as the Level of Service Inventory-Revised (LSI-R), the Self-Appraisal Questionnaire (SAQ), and the Historical-Clinical-Risk Management-20 (HCR-20), still guide assessors to consider static factors; however, by including potentially dynamic items, such as attitude and substance use, they may be sensitive to change in risk levels over time and can assist in identification of treatment targets. These tools are sometimes referred to as “risk-need” instruments and, unlike second generation assessments, tend to be theoretically- and empirically-based as opposed to wholly data driven.

### *Fourth Generation*

Most recently, fourth generation risk assessments explicitly integrate case planning and risk management into the assessment process. As such, the primary goal of the fourth generation extends beyond assessing risk and focuses on enhancing treatment and supervision. Examples of fourth generation tools include the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), Ohio Risk Assessment System (ORAS), and Wisconsin Risk and Needs tool (WRN). Like the third generation, this generation of risk assessment instruments allows for the role of professional judgment while remaining grounded in research and theory.

## ***Structured Approaches to Conducting Risk Assessments***

There are two broad categories that distinguish between the structured approaches used to conduct risk assessment in the second, third and fourth generations: actuarial and structured professional judgment. We briefly review the strengths and limitations of each below.

### *Actuarial Risk Assessment*

The actuarial approach represents a mechanical model of risk assessment, largely focused on historical or unchanging risk factors. When an actuarial instrument is used to assess risk, an offender is scored on a series of items that were most strongly associated with recidivism in the development sample. The offender’s total score is cross-referenced with an actuarial table that translates the score into an estimate of risk over a specified timeframe (e.g., 10 years). This estimate represents the percentage of participants in the instrument’s development study who received that score and recidivated. For example, if an offender receives a score of +5 on an instrument which is translated into a risk estimate of 60% over 10 years, this means that 60% of those individuals who received a score of +5 in the instrument’s original study went on to recidivate within that time. This does not mean that the offender has a 60% chance of recidivating over a period of 10 years. This is an important distinction that is frequently overlooked in practice.

Strengths of the actuarial approach include:

- *Objectivity.* No human judgment is involved in estimating risk once items have been rated. Items are typically straightforward and easy to rate (e.g., age, sex, number of prior offenses).
- *Accuracy.* Actuarial assessments are more accurate than unstructured assessments.
- *Transparency.* Information used to inform risk estimates is explicitly included in the instrument. Items are weighted in a pre-determined manner to compute total scores and estimate risk.
- *Speed.* Items included in actuarial instruments can usually be scored using information available in official records.

Drawbacks include the application of group-based statistics and norms to individual offenders. Beyond potential statistical issues (see Hart, Michie & Cooke, 2007), this is a concern because we do not know where any given offender falls within a risk bin. Using the same example provided earlier, if 60% of the individuals who received a score of +5 recidivated over a 10-year period, then 40% did not. Actuarial assessments cannot help distinguish whether an offender receiving a score of +5 is among the 60% or 40%. Additionally, with invariant item content comes the potential exclusion of case specific factors that do not systematically increase (or decrease) recidivism risk across the population but are relevant to a particular offender's level of risk. Finally, actuarial assessments speak to level of risk and may inform decisions regarding risk classification and allocation of resources. However, their utility in guiding the development and implementation of individualized risk reduction and rehabilitation plans is limited due to their focus largely on historical or unchangeable factors that cannot be addressed in treatment.

### *Structured Professional Judgment*

In contrast to the mechanistic, actuarial approach, the structured professional judgment approach focuses on creating individualized and coherent risk formulations and comprehensive risk management plans. These instruments act as *aide-mémoires*, guiding assessors to estimate risk level (e.g., low, moderate or high) through consideration of a set number of factors that are empirically and theoretically associated with the outcome of interest. Although offenders are scored on individual items, total scores are not used to make the final judgments of risk. Instead, assessors consider the relevance of each item to the individual offender, as well as whether there are any case specific factors not explicitly included in the list.

Strengths of the structured professional judgment approach include:

- *Professional discretion.* Assessors consider the relevance of factors to the individual offender to inform final estimates of each. Case specific factors also can be taken into consideration.
- *Accuracy.* Structured professional judgment assessments are more accurate than unstructured assessments (and comparable in accuracy to actuarial assessments).

- *Transparency.* Assessors rate a known list of factors according to specific guidelines. Additional items considered are added to the assessment form.
- *Risk communication and reduction.* Risk formulations provide information regarding the anticipated series of stressors and events that lead to the adverse outcome and over what period time, which can inform risk management strategies and identify treatment targets.

Drawbacks include the potential re-introduction of decision-making biases in the final risk judgments. Structured professional judgment instruments also take comparatively longer to administer than actuarial assessments; item ratings often are more nuanced and information might not be readily available on file to code all items. That said, recent reviews show that actuarial and structured professional judgment instruments produce assessments with commensurate rates of validity in predicting recidivism (Fazel, Singh, Doll & Grann, 2012).

### ***Types of Items and Content Domains***

Risk assessment instruments include items that represent characteristics of the offender (e.g., physical health, mental health, attitudes), his or her physical and/or social environment (e.g., neighborhood, family, peers) or circumstances (e.g., living situation, employment status) that are associated with the likelihood of offending. *Risk factors* are those characteristics that increase risk of offending, whereas *protective factors* are those that reduce risk. Inclusion of protective factors in risk assessment instruments—designed to assess recidivism risk or otherwise—is relatively rare; however, there is mounting evidence that they contribute unique information and improve predictive validity above and beyond consideration risk factors (e.g., Desmarais, Nicholls, Wilson, & Brink, 2012).

Most frequently, recidivism risk assessment instruments focus on biological, psychological and social characteristics; however, more macro-level factors—such as service, system and societal variables—also may affect risk, but are rarely included in recidivism risk assessment instruments.

In a relatively recent review of the literature, Andrews, Bonta and Wormith (2006) identified a shortlist of the most “powerful” risk factors for recidivism across offenders and situations. These include:

1. History of antisocial behavior
2. Antisocial personality pattern
3. Antisocial cognition
4. Antisocial associates
5. Family and/or marital problems
6. School and/or work problems
7. Leisure and/or recreation problems
8. Substance abuse

These “Central Eight” have been widely accepted as the most important domains to be assessed and targeted in risk assessment and management efforts.

Finally, risk and protective factors can either be static or dynamic in nature. *Static factors* are historical or otherwise unchangeable characteristics (e.g., history of antisocial behavior) that help establish absolute level of risk. In contrast, *dynamic factors* are changeable characteristics (e.g., substance abuse) that establish a relative level of risk and help inform intervention; they can be either relatively *stable*, changing relatively slowly over time (e.g., antisocial cognition) or *acute* (e.g., mood state) (Hanson & Harris, 2000). Research shows that dynamic factors add incrementally to the predictive validity of static factors and that the former may be more relevant to short-term outcomes and rehabilitation efforts (Wilson, Desmarais, Nicholls, Hart, & Brink, in press), whereas the latter to longer term outcomes and risk classification (Hart, Webster, & Douglas, 2001). Thus, there are important benefits to considering both static and dynamic factors in assessing recidivism risk.

### ***Focus of the Present Review***

In this review, we summarize the research conducted in the U.S. examining the performance of instruments designed to assess risk of recidivism among adult offenders, including new offenses and violation of conditions. We focus specifically on performance of tools validated and currently used in correctional settings in the United States.<sup>1</sup> By identifying those instruments that produce the most consistent and accurate assessments, decision makers may be able to make more informed choices regarding which measure(s) to implement and how they should invest financial and staff resources.

---

<sup>1</sup> For meta-analytic reviews of instruments used in other jurisdictions and research outside the United States see Fazel et al., 2012; Gendreau, Goggin, & Little, 1996; Smith, Cullen, & Latessa, 2009).



## METHODS OF THE CURRENT REVIEW

### *Search Criteria and Process*

#### *Identifying Risk Assessment Instruments Used in Correctional Settings in the U.S.*

Instruments designed to assess risk of recidivism were identified by searching academic research databases (PsycINFO and the U.S. National Criminal Justice Reference Service Abstracts) and Google using combinations of the following keywords: *risk assessment, instrument, tool, general, recidivism, offending, probation revocation, parole violation, and prediction*. We identified additional instruments by looking through the reference lists of recent publications and through discussion with colleagues.

We identified instruments designed to assess risk of recidivism by searching academic research databases and the Google search engine. We identified additional instruments by looking through the reference lists of recent publications and through discussion with colleagues. Criteria for instruments to be included in the review were: a) designed to assess the likelihood of general recidivism (i.e., new offenses and violation of conditions); b) intended for assessing adult offenders (18 years of age and older); c) currently or recently used in correctional settings in the United States; and d) validated in the United States.

Instruments were excluded from our review if they: a) were designed to assess the likelihood of specific offenses (e.g., sexual offenses, violent offenses, spousal assault); b) were intended for assessing juvenile offenders (less than 18 years of age); c) were not used in correctional settings in the United States; d) had not been validated in the United States; or e) were developed for use in a specific institution or ward.

We also excluded violence risk assessment instruments (e.g., Historical, Clinical, Risk Management-20, Violence Risk Appraisal Guide), clinical inventories (e.g., Beck Depression Inventory, Novaco Anger Scale), personality assessments (e.g., Psychopathy Checklist-Revised, Personality Assessment Inventory), and criminal thinking scales (e.g., TCU Criminal Thinking Scales, Psychological Inventory of Criminal Thinking) from our formal review. These instruments were not designed to assess risk for general offending *per se*; however, they frequently are used for that purpose in correctional settings in the U.S. Thus, we briefly review their validity in predicting general offending later in this report.

Using these inclusion and exclusion criteria, we identified 19 instruments:

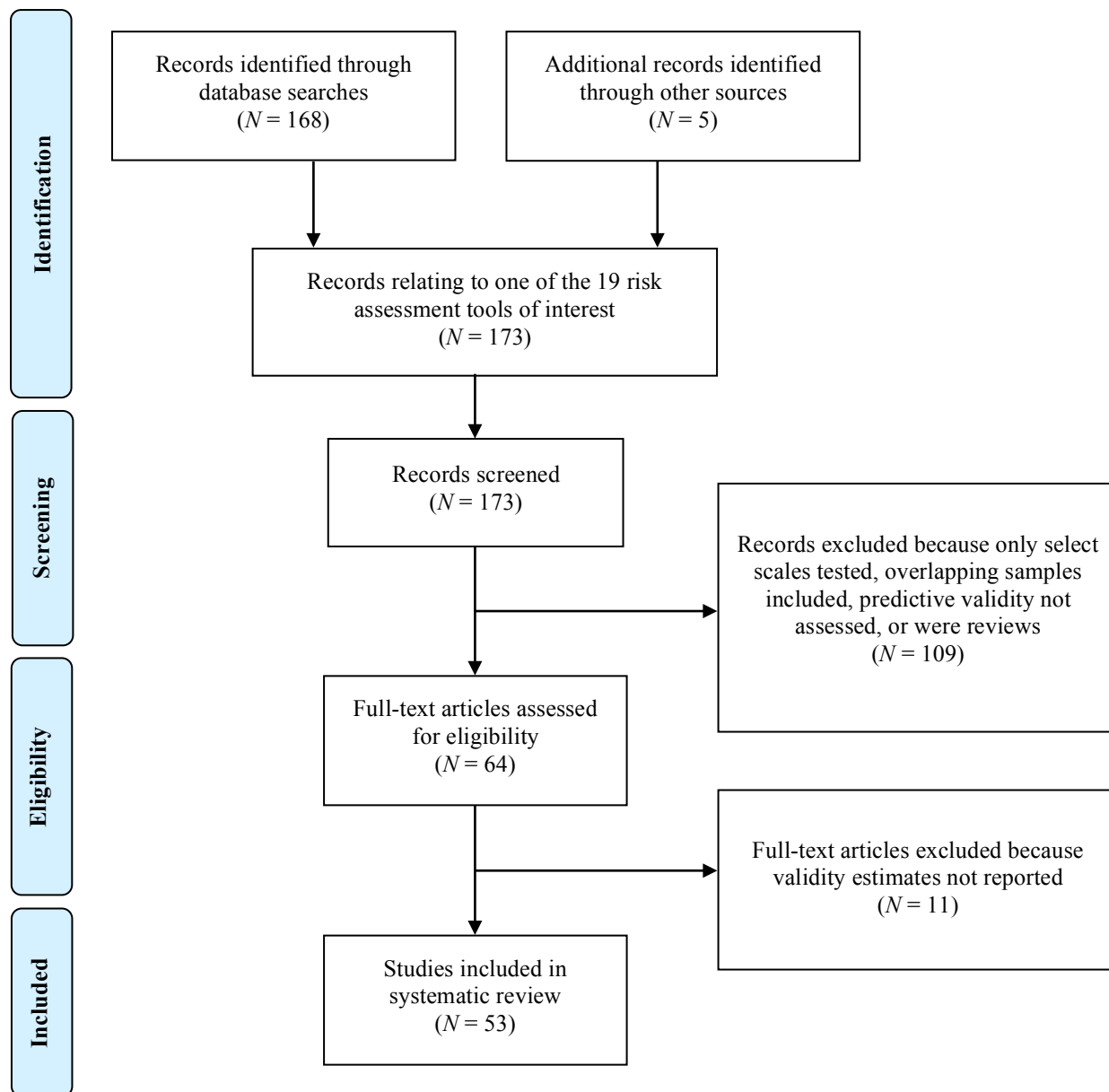
1. Community Risk/Needs Management Scale (CRNMS)
2. Correctional Assessment and Intervention System (CAIS)
3. Correctional Offender Management Profile for Alternative Sanctions (COMPAS)
4. Dynamic Factors Intake Assessment (DFIA)
5. Inventory of Offender Risks, Needs, and Strengths (IORNS)

6. Level of Service instruments, including Level of Service/Case Management Inventory (LS/CMI), Level of Service/Risk, Need, Responsivity (LS/RNR), Level of Service Inventory (LSI), Level of Service Inventory-Revised (LSI-R), and Level of Service Inventory-Revised: Screening Version (LSI-R:SV)
7. Offender Assessment System (OASys)
8. Offender Group Reconviction Scale (OGRS)
9. Ohio Risk Assessment System, including the Ohio Risk Assessment System-Pretrial Assessment Tool (ORAS-PAT), Ohio Risk Assessment System-Community Supervision Tool (ORAS-CST), Ohio Risk Assessment System-Community Supervision Screening Tool (ORAS- CSST), Ohio Risk Assessment System-Prison Intake Tool (ORAS-PIT), and Ohio Risk Assessment System-Reentry Tool (ORAS-RT)
10. Federal Post Conviction Risk Assessment (PCRA)
11. Recidivism Risk Assessment Scales (RISc)
12. Risk Management System (RMS)
13. Risk of Reconviction (ROC)
14. Statistical Information of Recidivism Scale (SIR)
15. Salient Factor Score instruments, including the Salient Factor Score-1974 Version (SFS74), Salient Factor Score-1976 Version (SFS76), and Salient Factor Score-1998 Version (SFS98)
16. Self-Appraisal Questionnaire (SAQ)
17. Service Planning Instrument (SPIn) and Service Planning Instrument-Women (SPIn-W)
18. Static Risk and Offender Needs Guide (STRONG)
19. Wisconsin Risk and Needs (WRN) and Wisconsin Risk and Needs-Revised (WRN-R)

We also identified 47 instruments designed for use in specific jurisdictions. Detailed review is beyond the scope of the current report, but these instruments are listed in Appendix A.

### *Identifying Predictive Validity Studies*

Studies investigating the predictive validity of the 19 above instruments were identified using the same databases, search engine and secondary sources as above, using both the acronyms and full names of the instruments as search criteria. We searched for studies published between 1970 and 2012 in peer-reviewed journals, as well as government reports, doctoral dissertations, and Master's theses. Studies were included in our review if their titles, abstracts, or methods sections described evaluations of validity in predicting general offending (including the violation of probation or parole conditions) conducted in the U.S. Studies were excluded if they only included some items or scales of an instrument. Using this search strategy, an initial total of 173 records was filtered to a final count of 53 studies ( $k$  samples = 72), including 26 journal articles ( $k = 30$ ), 16 government reports ( $k = 31$ ), nine doctoral dissertations ( $k = 9$ ), and two Master's theses ( $k = 2$ ). This systematic search process is visually depicted in the figure on the following page. A full list of the included studies is available from the authors upon request.

**Systematic Search Conducted to Identify U.S. Predictive Validity Studies**

### *Evaluation Criteria and Process*

Three research assistants collected information about the characteristics of the risk assessment instruments (approach, number of items, types of items, domains measured, intended population and outcome) and studies (geographic location, context, design, population, sample size, sex, race/ethnicity, age, diagnostic composition, outcome, length of follow-up), as well as characteristics of the assessment process (setting, timing, format, assessor, sources of information, time needed to administer and score) from the included studies. They recorded information on inter-rater reliability and predictive validity, overall and by offender sex, race/ethnicity, study context, and recidivism outcome, where possible.

To evaluate performance, we computed the median performance indicators reported across studies for inter-rater reliability and predictive validity. For inter-rater reliability, we used the criteria presented in Table 1 to determine the practical significance of the median indicators.

**Table 1. Criteria Used to Determine Practical Significance of Aggregate Inter-Rater Reliability Findings**

INTER-RATER RELIABILITY	PERFORMANCE INDICATOR		
	Kappa ( $\kappa$ )	Intra-class Correlation Coefficient (ICC)	Observed Agreement (%)
<b>Poor</b>	.00 – .40	.00 – .40	< 70
<b>Fair</b>	.40 – .59	.40 – .59	70 – 79
<b>Good</b>	.60 – .74	.60 – .74	80 – 89
<b>Excellent</b>	.75 – 1.00	.75 – 1.00	90 – 100

*Note.* Table adapted from Cicchetti (2001, p. 697).

We also computed the median performance indicators for predictive validity. We used the criteria presented in Table 2 to determine the practical significance.

**Table 2. Criteria Used to Determine Practical Significance of Aggregate Predictive Validity Findings**

PREDICTIVE VALIDITY	PERFORMANCE INDICATOR				
	Cohen's <i>d</i>	Correlation ( $r_{pb}$ )	Area Under the Curve (AUC)	Odds Ratio (OR)	Somer's <i>d</i>
<b>Poor</b>	< .20	< .10	< .55	< 1.50	< .10
<b>Fair</b>	.20 – .49	.10 – .23	.55 – .63	1.50 – 2.99	.10 – .19
<b>Good</b>	.50 – .79	.24 – .36	.64 – .71	3.00 – 4.99	.20 – .29
<b>Excellent</b>	$\geq$ .80	.37 – 1.00	.71 – 1.00	$\geq$ 5.00	.30 – 1.00

*Notes.* Criteria were anchored to Cohen's *d* (1988) and based upon the calculations of Rice and Harris (2005) for AUC values, and Chen, Cohen, and Chen (2010) for the odds ratios. Somer's *d* values, as well as those for other performance indicators reported less frequently, also were interpreted in relation to Cohen's *d*.

In following sections of this report, we first summarize findings across instruments and then present findings of this review by instrument, respectively. We report only the interpretations of the practical significance of the performance indicators for both inter-rater reliability and predictive validity, but detailed statistical results are available upon request. We did not find any studies investigating the predictive validity of the CAIS, CRNMS, DFIA, LS/CMI, LS/RNR, LSI, OGRS, OASys, RISc, ROC, SFS98, SIR, or SPIn that met our inclusion criteria.

For a glossary of terms used in this report, including a brief explanation of the performance indicators included in Tables 1 and 2, see Appendix B.

## SUMMARY OF FINDINGS ACROSS INSTRUMENTS

### *Characteristics of the Risk Assessment Instruments*

Table 3 summarizes the characteristics of the risk assessment instruments. The number of items ranged from four for the ORAS-CSST to 130 for the IORNS. All instruments were intended for use across offender populations, with the exception of the SFS74, SFS76 and SFS81. Most were intended to be used to assess risk of new offenses, excluding violations). Of the nine instruments for which estimates were provided in the manual, length ranged from 5-10 minutes for the ORAS-CSST up to 60 minutes for the COMPAS. All were actuarial instruments.

**Table 3. Characteristics of Risk Assessment Instruments**

INSTRUMENTS	CHARACTERISTICS					
	<i>k</i>	Items	Generation	Intended Population(s)	Intended Outcome(s)	Time (minutes)
COMPAS	3	70	4 <sup>th</sup>	Any Offender	Offenses & Violations	10-60
IORNS	1	130	3 <sup>rd</sup>	Any Offender	Offenses & Violations	15-20
LSI-R	25	54	3 <sup>rd</sup>	Any Offender	Offenses & Violations	30-40
LSI-R:SV	2	8	3 <sup>rd</sup>	Any Offender	Offenses & Violations	10-15
ORAS-PAT	3	7	4 <sup>th</sup>	Any Offender	Offenses	10-15
ORAS-CST	1	35	4 <sup>th</sup>	Any Offender	Offenses	30-45
ORAS-CSST	1	4	4 <sup>th</sup>	Any Offender	Offenses	5-10
ORAS-PIT	1	31	4 <sup>th</sup>	Any Offender	Offenses	Unknown
ORAS-RT	1	20	4 <sup>th</sup>	Any Offender	Offenses	Unknown
PCRA	2	30	4 <sup>th</sup>	Any Offender	Offenses & Violations	15-30
RMS	2	65	4 <sup>th*</sup>	Any Offender	Offenses	Unknown
SAQ	2	72	3 <sup>rd</sup>	Any Offender	Offenses	15
SFS74	3	9	2 <sup>nd</sup>	Parolees	Offenses	Unknown
SFS76	4	7	2 <sup>nd</sup>	Parolees	Offenses	Unknown
SFS81	8	6	2 <sup>nd</sup>	Parolees	Offenses	Unknown
SPIn-W	2	100	4 <sup>th</sup>	Any Offender	Offenses	Unknown
STRONG <sup>a</sup>	1	26	4 <sup>th</sup>	Any Offender	Offenses	Unknown
WRN	9	53	4 <sup>th</sup>	Any Offender	Offenses	Unknown
WRN-R	1	52	4 <sup>th</sup>	Any Offender	Offenses	Unknown

*Notes.* *k* = number of samples; Offenses = new arrest, charge, conviction, or incarceration; Violations = violations of conditions of supervision. <sup>a</sup>The STRONG includes three parts; table values reflect only the first part, which is the component used to assess risk of recidivism. \*The authors of the RMS describe it as being a 5<sup>th</sup> generation risk assessment instrument due to its exemplar-based approach.

Table 4 summarizes the types of factors included in the instruments. Only two instruments, the IORNS and the SPIn-W, include protective factors; all others include risk factors exclusively. The majority include static and dynamic factors, with the exception of the SFS instruments and the STRONG, both of which only include static factors. None only include only dynamic factors.

**Table 4. Types of Items Included in the Risk Assessment Instruments**

INSTRUMENTS	TYPES OF ITEMS			
	Risk	Protective	Static	Dynamic
COMPAS	X		X	X
IORNS	X	X	X	X
LSI-R	X		X	X
LSI-R:SV	X		X	X
ORAS-PAT	X		X	X
ORAS-CST	X		X	X
ORAS-CSST	X		X	X
ORAS-PIT	X		X	X
ORAS-RT	X		X	X
PCRA	X		X	X
RMS	X		X	X
SAQ	X		X	X
SFS74	X		X	
SFS76	X		X	
SFS81	X		X	
SPIn-W	X	X	X	X
STRONG <sup>a</sup>	X		X	
WRN	X		X	X
WRN-R	X		X	X

*Note.* <sup>a</sup>The STRONG includes three parts; table values reflect only the first part, which is the component used to assess risk of recidivism.

Table 5 summarizes the content domains considered in the risk assessment instruments. All instruments include items assessing history of antisocial behavior and substance use problems. Slightly more than half of the instruments have items assessing mental health problems. Nine instruments include items assessing personality problems. Roughly two-thirds of the instruments consider attitudes, and similar proportions consider the influence of peers and relationships. The COMPAS and the LSI-R consider the most content domains. The ORAS-CST, ORAS-PIT, RMS, and SPIn-W evaluate all but one of the domains included in Table 5; the exception varied for each instrument. The SFS81 and STRONG instruments considered the fewest domains.

Table 5. Content Domains Assessed across the Risk Assessment Instruments

INSTRUMENTS	ITEM CONTENT DOMAINS									
	Attitudes	Associates/ Peers	History of Antisocial Behavior	Personality Problems	Relationships	Work/ School	Recreation/ Leisure Activities	Substance Use Problems	Mental Health Problems	Housing Status
COMPAS	X	X	X	X	X	X	X	X	X	X
IORNS	X	X	X	X	X	X		X	X	
LSI-R	X	X	X	X	X	X	X	X	X	X
LSI-R:SV	X	X	X		X	X		X	X	
ORAS-PAT			X			X		X		X
ORAS-CST	X	X	X	X	X	X	X	X		X
ORAS-CSST		X	X			X		X		
ORAS-PIT		X	X	X	X	X	X	X	X	X
ORAS-RT	X		X	X	X	X		X	X	
PCRA	X	X	X		X	X		X		
RMS	X	X	X	X	X	X		X	X	X
SAQ	X	X	X	X				X		
SFS74			X			X		X		X
SFS76			X			X		X		
SFS81			X					X		
SPIIn-W	X	X	X		X	X	X	X	X	X
STRONG			X					X		
WRN	X	X	X		X	X		X	X	
WRN-R	X	X	X		X	X		X	X	

Note. <sup>a</sup>The STRONG includes three parts; table values reflect only the first part, which is the component used to assess risk of recidivism.



## ***Study Characteristics***

### *Population and Sample Characteristics*

More than a third of samples (40%) comprised inmates and roughly a quarter (22%), probationers. The remainder included at either parolees only (11%) or inmates and parolees (7%) or probationers and parolees (11%). Legal status was not reported in six samples (8%).

Studies generally provided few details regarding sample characteristics. Below we summarize findings regarding size, age, sex, race/ethnicity and mental health, when reported.

*Sample size.* The average sample size after attrition was 5,032.

*Age.* The average offender age at the time of risk assessment was 33.5 years.

*Sex.* In samples where sex was reported, the vast majority of offenders (86%) were male.

*Race/ethnicity.* In samples where race/ethnicity was reported, almost two-thirds (61%) were White and close to one-third (29%) were Black, with 14% identified as Hispanic. It is important to note that racial/ethnic categories were not consistent across studies. For instance, in some cases, authors reported the proportion of offenders identified as White, Black, or Hispanic (Farabee et al., 2010), while others reported prevalence of Hispanic and non-Hispanic offenders (Tillyer & Vose, 2011).

*Mental health.* Mental health characteristics were rarely reported. Only five studies--one evaluating the SFS74, one evaluating the SFS81, two evaluating the SPIn-W and one evaluating the WRN--described prevalence of major mental disorder (MMD), substance use disorder (SUD), or personality disorder. All offenders in the Howard (2007) study of the SFS81 were diagnosed with an MMD; slightly under half (46%) an SUD, and 11% had a personality disorder. This was the only study reporting prevalence of personality disorders. In one study of the SPIn-W all offenders had an SUD and three-quarters, a MMD (Meadon, 2012), whereas in the other study of the SPIn-W, just over half (53%) had a MMD (Millson et al., 2010). Only the WRN study reported prevalence by diagnosis. Bipolar disorder was the most prevalent MMD (36%) and schizophrenia, the least (16%), and alcohol abuse was the most prevalent SUD (48%) and amphetamines, the least (13%) (Castillo & Alardi, 2011). Finally, in the SFS74 study (Robuck, 1976), just under half of the sample (47%) suffered from alcohol abuse and 15%, illicit drug use.

### *Assessment Process*

Table 6 shows the characteristics of the assessment process used in the studies. Risk assessments were complete by professionals in forensic services for over three-quarters of the studies (82%); the remaining assessments were conducted by the researchers (15%) or, in two studies, were self-administered. These assessments most often took place in a prison (28%) or in the community (38%), but at times were administered in jail (10%), a clinic or hospital (4%), or at another facility (6%). In terms of timing, roughly one third of assessments (36%) were conducted during community supervision, a quarter were completed pre-release (26%), and the remainder were conducted either prior to incarceration (11%) or at admission (10%). The source of information

used to complete the assessments were file reviews in 24 samples (33%), interviews in 12 samples (17%), and offender self-report in two samples (3%).

**Table 6. Characteristics of the Assessment Process Used in Studies Included in this Review**

CHARACTERISTICS	NUMBER OF SAMPLES
	<i>k</i> (%)
<b>Assessor</b>	
Researcher	11 (15.3)
Professional	59 (81.9) <sup>a</sup>
Offender (self-report)	2 (2.8) <sup>b</sup>
<b>Assessment Setting</b>	
Jail	7 (9.7)
Prison	20 (27.8)
Clinic/Hospital	3 (4.2)
Community	27 (37.5)
Other	4 (5.6)
Unstated/Unclear	11 (15.3)
<b>Timing of Assessment</b>	
Prior to incarceration	8 (11.1)
At admission	7 (9.7)
Prior to release	19 (26.4)
During community supervision	26 (36.1)
Unstated/Unclear	13 (18.1)
<b>Source(s) of Information</b>	
File review	24 (33.3)
Interview	12 (16.7)
Self-report	2 (2.8)
Mixed	18 (25.0)
Unstated/Unclear	16 (22.2)

*Notes.* Overall  $k = 72$  samples. <sup>a</sup>Correctional officer ( $k = 35$ , 48.6%), parole officer ( $k = 2$ , 2.8%), probation officer ( $k = 1$ , 1.4%), other trained staff ( $k = 14$ , 19.4%), unstated/unclear ( $k = 7$ , 9.7%). <sup>b</sup>The SAQ, designed to be self-administered, was the only tool not administered by a researcher or professional.

Administration time was reported for only five instruments in a total of nine studies. For the LSI-R administration time ranged from 30 to 60 minutes for assessments conducted in the context of ‘real world’ practice (Holsinger et al., 2004; Lowenkamp et al., 2009), and 45 to 90 minutes in research studies (Evans, 2009; Latessa et al., 2009). The LSI-R:SV was reported to have a mean administration time of 10 minutes when completed in practice (Miller, 2006). In the same study, the IORNS required 15 minutes to complete; however, this estimate included only the interview portion of the assessment. Across three studies, administration time for the COMPAS varied

from 43 to 165 minutes (Brennan et al., 2009; Farabee et al., 2010; Farabee & Zhang, 2007). In the study evaluating SAQ assessments, assessments were reported to take approximately 20 minutes (Mitchell & McKenzie, 2006).

### *Study Designs and Procedures*

More than two-thirds of studies (70%) used a prospective study design, an optimal approach for examining predictive validity, and the average length of follow-up was almost two years (23.5 months). Studies were most frequently conducted in midwestern states (38%) followed by the southwestern and northeastern (11% each) regions of the U.S.

Close to 70% of the studies examined general recidivism as the outcome; roughly a quarter (26%) considered a variety of outcomes, and the remainder (18%) focused specifically on violations. As a result, our knowledge of the validity of recidivism risk assessment instruments in predicting violations as opposed to other forms of recidivism is limited. The threshold for recidivism varied across studies, but arrest was used as an indicator in close to a third of studies (31%), followed in order by conviction (13%), incarceration (10%), revocations (4%), and charge (3%). Finally, assessments for the majority of samples (65%) were conducted in the context of 'real world' practice rather than for the purposes of research.

Nearly a third of the studies included in our review (31%,  $k = 22$ ) were conducted by the author of the tool being studied. In fact, for many instruments, all of the studies included in our review were completed by the same people who developed the instrument under investigation. This was true for the IORNS (Miller, 2006), the PCRA (Johnson et al., 2011), the ORAS instruments (Latessa et al., 2008, 2009), the STRONG (Barnoski & Drake, 2007), and the WRN-R (Eisenberg et al., 2009). The authors of the RMS conducted one of two studies evaluating predictive validity of RMS assessments (Dow et al., 2005), and the authors of the COMPAS conducted one of three samples evaluating COMPAS assessments (Brennan et al., 2009). The authors of the SFS74, SFS76, and SFS81 evaluated two of three samples for the SFS74 (Hoffman & Beck, 1974), two of four for the SFS76 (Hoffman, 1980; Hoffman & Beck, 1980), and four of eight for the SFS81 (Hoffman, 1983, 1994; Hoffman & Beck, 1985).

**Table 7. Design Characteristics and Procedures of Studies Included in this Review**

CHARACTERISTICS	NUMBER OF SAMPLES
	<i>k</i> (%)
<b>Study Context</b>	
Research	25 (34.7)
Practice	47 (65.3)
<b>Temporal Design</b>	
Prospective	50 (69.4)
Retrospective	22 (30.6)
<b>Geographical Region</b>	
Northwest	2 (2.8)
Southwest	8 (11.1)
Midwest	27 (37.5)
Northeast	8 (11.1)
Southeast	5 (6.9)
Non-continental	1 (1.4)
Mixture	1 (1.4)
Unstated/Unclear	20 (27.8)
<b>Type of Outcome</b>	
General recidivism	50 (69.4)
Violation/Breach of conditions	13 (18.1)
Mixed	19 (26.4)
<b>Threshold for Recidivism</b>	
Arrest	22 (30.6)
Charge	2 (2.8)
Conviction	9 (12.5)
Incarceration	7 (9.7)
Revocation	3 (4.2)
Mixed	29 (40.3)

*Note.* *k* = number of samples

### ***Inter-Rater Reliability***

Inter-rater reliability was evaluated in only two studies, one examining the LSI-R and the other, the LSI-R:SV. In both cases, inter-rater reliability was excellent. Assessments were conducted by professionals rather than research assistants, providing evidence of *field* reliability, specifically.

## ***Predictive Validity***

### *Overall*

Table 8 presents the practical significance of predictive validity performance indicators across studies. Overall, and consistent with prior research reviews, no one instrument stands out as producing more accurate instruments than the others, with validity varying with the indicator reported. Odds ratios generally suggested poor performance for the majority of instruments, with only one instrument (the SFS81) demonstrating good predictive validity. In contrast, Somer's *d* values ranged from good to excellent. AUCs and point-biserial correlations each ranged from fair to excellent across instruments. Below, we describe predictive validity by instrument.

*COMPAS.* The predictive validity of COMPAS assessments ranged from poor to good, as a function of performance indicator; more studies used the AUC and, thus, reported good validity.

*LSI instruments.* LSI-R assessments were evaluated in the most samples. Predictive validity was good across studies and indicators, with the exception of odds ratios. Validity of LSI-R:SV assessments ranged from fair to good.

*ORAS instruments.* Across instruments and studies, ORAS assessments demonstrated excellent point-biserial values. No other performance indicators were reported.

*PCRA.* PCRA assessments were evaluated in only two samples, with AUC values suggesting excellent predictive validity in both. No other performance indicators were reported.

*RMS.* In three samples, RMS assessments showed good performance according to the AUC values. No other performance indicators were reported.

*SFS instruments.* SFS74, SFS76, and SFS81 assessments showed predictive validity ranging from good to excellent, with the SFS81 outperforming the previous versions.

*SPIn-W.* SPIn-W assessments showed good performance according to the AUC but poor performance according to the odds ratio.

*STRONG.* In one study, predictive validity of STRONG assessments was excellent according to the AUC. No other performance indicators were reported.

*WRN instruments.* Predictive validity for WRN and WRN-R assessments ranged from poor to good, depending on the performance indicator used.

No studies reported predictive validity of IORNS or SAQ assessments using these indicators.

**Table 8. Summary of Predictive Validity Findings by Performance Indicator across Studies**

INSTRUMENT	MEDIAN PERFORMANCE INDICATOR							
	<i>k</i>	AUC	<i>K</i>	<i>r</i> <sub>pb</sub>	<i>k</i>	OR	<i>k</i>	Somer's <i>d</i>
COMPAS	3	Good	1	Fair	1	Poor	–	–
LSI-R	5	Good	21	Good	6	Poor	2	Good
LSI-R:SV	1	Fair	1	Good	–	–	–	–
ORAS-PAT	–	–	5	Good	–	–	–	–
ORAS-CST	–	–	1	Excellent	–	–	–	–
ORAS-CSST	–	–	1	Excellent	–	–	–	–
ORAS-PIT	–	–	1	Excellent	–	–	–	–
ORAS-RT	–	–	1	Excellent	–	–	–	–
PCRA	2	Excellent	–	–	–	–	–	–
RMS	3	Good	–	–	–	–	–	–
SFS74	–	–	–	–	–	–	2	Good
SFS76	–	–	1	Excellent	–	–	2	Good
SFS81	–	–	4	Excellent	2	Good	5	Excellent
SPIn-W	1	Excellent	–	–	1	Poor	–	–
STRONG	1	Excellent	–	–	–	–	–	–
WRN	3	Good	6	Fair	1	Poor	–	–
WRN-R	1	Good	–	–	–	–	–	–

*Notes.* *k* = number of samples; AUC = area under the receiver operating characteristic curve; *r*<sub>pb</sub> = point-biserial correlation coefficient; OR = odds ratio. Medians were calculated using either total scores or risk bins. There were no studies reporting predictive validity of the IORNS or SAQ using these performance indicators.

### *Validity of Total Scores in Predicting Different Forms of Recidivism*

Table 9 presents the validity of total scores in predicting different forms of recidivism. For general offending *including* violations, predictive validity ranged from poor for SPIn-W assessments to excellent for SFS76 and SFS81 assessments. For general offending *excluding* violations, total scores for over two-thirds of instruments had either good or excellent predictive validity. Specifically, predictive validity ranged from fair for ORAS-PAT assessments to excellent for the ORAS-CST, ORAS-CSST, PCRA, and STRONG assessments. For *violations*, predictive validity ranged from fair COMPAS assessments to excellent WRN assessments. Below, we describe predictive validity by instrument.

*COMPAS.* The COMPAS total scores demonstrated good validity in predicting general offending *excluding* violations, but was only fair for violations only.

*LSI instruments.* LSI-R total scores showed good predictive validity for both general offending *including* violations and violations only, and ranged from fair to good validity in general offending *excluding* violations.

*ORAS instruments.* With the exception of the ORAS-PAT, the total scores on the ORAS instruments all demonstrated predictive validity ranging from good to excellent for general offending *excluding* violations. ORAS-PAT total scores, however, were only fair at predicting general offending outcomes, though predictive validity was good for violations only.

*RMS.* RMS total scores demonstrated good validity in predicting general offending *excluding* violations, as well as violations only.

*SFS instruments.* SFS76 and SFS81 total scores showed excellent validity in predicting general offending *including* violations. No studies reported predictive validity of SFS74 total scores by outcome.

*SPIn-W.* SPIn-W total scores had poor validity in predicting general offending *including* violations.

*STRONG.* STRONG total scores demonstrated excellent validity in predicting general offending *excluding* violations.

*WRN instruments.* WRN total scores ranged from fair to good in their ability to predict general offending *excluding* violations. Predictive validity was excellent for violations only. WRN-R total scores showed good validity in predicting general offending *excluding* violations.

Overall, total scores of SFS76 and SFS81 total scores stood out as excellent predictors of general offending *including* violations. Total scores on the ORAS-CST, ORAS-CSST, PCRA, and STRONG were excellent predictors of general offending *excluding* violations. WRN total scores stood alone as excellent in predicting violations only. It is important to note, however, the small number of studies examining these outcomes; SFS76, ORAS-CST, ORAS-CSST, STRONG, and WRN assessments were evaluated in only one sample, compared to the 26 samples evaluating LSI-R assessments.

**Table 9. Validity of Total Scores in Predicting Different Forms of Recidivism**

INSTRUMENTS	OUTCOMES					
	<i>k</i>	General Offending (including Violations)	<i>k</i>	General Offending (excluding Violations)	<i>k</i>	Violations Only
COMPAS	–	–	5	Good	1	Fair
LSI-R	3	Good	26	Fair-Good	7	Good
LSI-R:SV	–	–	2	Fair-Good	–	–
ORAS-PAT	1	Fair	2	Fair	2	Good
ORAS-CST	–	–	1	Excellent	–	–
ORAS-CSST	–	–	1	Excellent	–	–
ORAS-PIT	–	–	1	Good	–	–
ORAS-RT	–	–	1	Good	–	–
PCRA	–	–	2	Excellent	–	–
RMS	–	–	1	Good	1	Good
SFS74	–	–	–	–	–	–
SFS76	1	Excellent	–	–	–	–
SFS81	6	Excellent	–	–	–	–
SPIIn-W	1	Poor	–	–	–	–
STRONG	–	–	1	Excellent	–	–
WRN	–	–	8	Fair-Good	1	Excellent
WRN-R	–	–	1	Good	–	–

*Notes.* *k* = number of samples. General Offending = new arrest, charge, conviction, or incarceration; Violations = violations of conditions of supervision.

### *Predictive Validity of Risk Classifications*

Table 10 presents the validity of risk classifications in predicting different forms of recidivism. Validity of risk classifications in predicting general offending *including* violations was excellent for SFS74, SFS76, and SPIIn-W assessments. For general offending *excluding* violations, the predictive validity was fair for WRN assessments and excellent for RMS and SFS81 assessments. Validity of SFS risk classifications in predicting general offending *including* violations also was excellent.

No U.S. studies examined the predictive validity of risk classifications for violations alone. There also were no U.S. studies reporting predictive validity of the risk classifications for the COMPAS, IORNS, LSI-R, LSI-R:SV, ORAS-PAT, ORAS-CST, ORAS-CSST, ORAS-PIT, ORAS-RT, PCRA, SAQ, STRONG, or WRN-R for any of the recidivism outcomes.



**Table 10. Validity of Risk Classifications in Predicting Different Forms of Recidivism**

INSTRUMENTS	OUTCOMES			
	<i>k</i>	General Offending (including Violations)	<i>k</i>	General Offending (excluding Violations)
RMS	–	–	1	Excellent
SFS74	2	Excellent	–	–
SFS76	2	Excellent	–	–
SFS81	4	Excellent	1	Excellent
SPIIn-W	1	Excellent	–	–
WRN	–	–	1	Fair

*Notes.* *k* = number of samples. There were no studies that reported the predictive validity of the risk classifications for the COMPAS, IORNS, LSI-R, LSI-R:SV, ORAS-PAT, ORAS-CST, ORAS-CSST, ORAS-PIT, ORAS-RT, PCRA, SAQ, STRONG, or WRN-R using these performance indicators. The risk bins used to classify offenders were those recommended by instrument authors.

#### *Predictive Validity across Offender Subgroups*

*Sex.* Table 11 presents the validity of total scores in predicting recidivism by the offender's sex. Overall, predictive validity ranged from fair to excellent for both male and female offenders. Some instruments performed equally well for male and female offenders; for instance, COMPAS assessments demonstrated good predictive validity for both sexes. STRONG assessments also demonstrated excellent validity for both male and female offenders. Finally, predictive validity for the ORAS instrument for which comparisons were possible—namely, the ORAS-CST, ORAS-CSST, ORAS-PIT, and ORAS-RT—ranged from good to excellent for both male and female offenders.

Other instruments showed differential performance by offender sex. In particular, LSI-R assessments showed good predictive validity for male offenders, but predictive validity was only fair for female offenders. Similarly, LSI-R:SV assessments showed only fair predictive validity for female offenders, but ranged from fair to good in its predictions for male offenders.

Other instruments were evaluated in exclusively male or female offenders. Predictive validity of SFS76 and SFS81 assessments, for example, were only evaluated for male offenders; SFS76 total scores demonstrated excellent validity, while validity of SFS81 assessments ranged from good to excellent. WRN total scores also were evaluated for male offenders and showed fair validity. Designed for women, the SPIIn-W has only been evaluated for female offenders and showed good validity.

No studies reported predictive validity of assessments by offender sex for the IORNS, ORAS-PAT, PCRA, RMS, SAQ, SFS74, or WRN-R.

**Table 11. Validity of Total Scores in Predicting Recidivism by Offender Sex**

INSTRUMENTS	OFFENDER SEX			
	<i>k</i>	Male	<i>k</i>	Female
COMPAS	2	Good	2	Good
LSI-R <sup>a</sup>	9	Good	8	Fair
LSI-R:SV	2	Fair-Good	1	Fair
ORAS-CST	1	Excellent	1	Good
ORAS-CSST	1	Good	1	Excellent
ORAS-PIT	1	Good	1	Good
ORAS-RT	1	Good	1	Excellent
SFS76 <sup>b</sup>	1	Excellent	–	–
SFS81 <sup>c</sup>	–	Good-Excellent	–	–
SPIn-W <sup>d,e</sup>	–	–	2	Good
STRONG	1	Excellent	1	Excellent
WRN	1	Fair	–	–

*Notes.* *k* = number of performance indicators. No studies reported predictive validity estimates by sex for the IORNS, ORAS-PAT, PCRA, RMS, SAQ, SFS74, or WRN-R using the included performance indicators.

<sup>a</sup>One LSI-R sample specifically included technical violations in the operational definition of recidivism.

<sup>b</sup>One SFS76 sample specifically included technical violations in the operational definition of recidivism.

<sup>c</sup>One SFS81 sample specifically included technical violations in the operational definition of recidivism.

<sup>d</sup>Both SPIn-W samples were composed entirely of women.

<sup>e</sup>One SPIn-W sample reported predictive validity of the risk categorizations rather than total scores.

*Race/ethnicity.* Comparisons by offender race/ethnicity were only possible for assessments completed using the COMPAS and LSI-R. For COMPAS assessments, predictive validity was good for White and Black offenders. For LSI-R assessments, predictive validity ranged from poor to good across White, Black, Hispanic, and non-White offenders, with performance varying largely depending on sample size and performance indicator rather than race/ethnicity. Together, these findings fail to provide evidence of differential performance of COMPAS and LSI-R assessments as a function of offender race/ethnicity.

*Diagnostic categories.* No comparisons of predictive validity within or across instruments as a function of mental, substance use or personality disorders were possible. Even when these sample characteristics were reported, predictive validity was not provided by subgroup. As for race/ethnicity, there is a critical need for research examining risk assessment accuracy between mentally disordered and nondisordered offenders as well as across diagnostic subgroups. That said, prior meta-analytic work has found the predictors of recidivism to be comparable for mentally disordered offenders (Bonta, Law, & Hanson, 1998), suggesting that assessments also may perform comparably.

*Predictive Validity in the Context of Research versus 'Real World' Practice*

Recently there has been a focus on the need to establish the performance of risk assessment instruments *in the field*. Much of our knowledge stems from research-based studies, in which researchers can carefully train and monitor assessors. In 'real world' practice, however, such training and oversight is not necessarily present (Douglas, Otto, Desmarais, & Borum, in press).

Comparisons between the performance of assessments completed in the context of research and practice were possible for the LSI-R, RMS, SPIn-W, and WRN. Whereas both LSI-R and WRN total scores performed comparably whether conducted in research studies or in the context of 'real world' practice, RMS risk classifications had better predictive validity when completed by researchers rather than practitioners (though performance was still good). SPIn-W assessments also seemed to perform better in research studies than in practice, though predictive validity in both contexts was excellent. However, in the research context, predictive validity of the SPIn-W was evaluated vis-à-vis the total scores while in practice, the risk classifications were examined, preventing direct comparisons of the results.

No comparisons were possible for the other risk assessment instruments. Specifically, COMPAS, IORNS, SFS76, and SFS81 assessments have only been evaluated in the context of 'real world' practice, and the LSI-R:SV, ORAS tools, PCRA, SAQ, SFS74, STRONG, and WRN-R assessments have only been evaluated in research studies.

## SUMMARY OF FINDINGS BY INSTRUMENT

In this section describe each risk assessment instrument and summarize findings of U.S. studies examining predictive validity. Instruments are presented in alphabetical order.

### *Correctional Offender Management Profiling for Alternative Sanctions*

#### *Description*

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations (Brennan et al., 2009).

The COMPAS contains static and dynamic risk factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, circumstances at school or work, leisure or recreational activities, substance use problems, mental health problems, and housing, divided across 22 scales (Blomberg, Bales, Mann, Meldrum, & Nedelec, 2010). Scores on the self-report assessment, data from official records, and information from interview are used to arrive at an overall risk score for each offender. The COMPAS is a 4<sup>th</sup> generation risk assessment instrument.

COMPAS assessments are completed through a combination of a computer-assisted self-report questionnaire, an interview conducted by a trained assessor, and data collected from the offender's records. The instrument can be purchased from Northpointe at [www.northpointeinc.com](http://www.northpointeinc.com). Assessors must complete a 2-day training session that covers practical use, interpretation of results, and case planning strategies in order to administer the COMPAS. Advanced training options that focus on the theoretical underpinnings of offender assessments, gender responsiveness, motivational interviewing, and other topics are available.

#### *U.S. Research Evidence*

In total, four studies have evaluated predictive validity of COMPAS assessments in U.S. samples. Blomberg and colleagues (2010) found that those identified as higher risk were indeed more likely to recidivate; specifically, 7% of those identified to be low risk recidivated, 16% of those identified as medium risk, and 27% of those identified as high risk. In other samples, predictive validity was good for general offending (Brennan, Dieterich, & Ehret, 2009) and fair for violations (Farabee & Zhang, 2007). Predictive validity for male and female offenders has ranged from good to excellent (Brennan et al., 2009).

There were no studies published between 1970 and 2012 comparing predictive validity in U.S. samples between total scores and risk classifications, assessments completed in research and practice contexts, or by offender race/ethnicity. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

### *Practical Issues and Considerations*

For the self-report portion of the assessment, the computer upon which the offender completes the questionnaire must have Internet access and run on Windows. The assessor must complete training to be qualified to administer the structured interview.

### *Selected References and Suggested Readings*

Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). *Validation of the COMPAS risk assessment classification instrument*. City, ST: publisher. Retrieved from <http://www.criminologycenter.fsu.edu/p/pdf/pretrial/Broward%20Co.%20COMPAS%20Validation%202010.pdf>

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36, 21-40.

Farabee, D., Zhang, S., Roberts, R. E. L., & Yang, J. (2010). *COMPAS validation study: Final report*. California Department of Corrections and Rehabilitation. Retrieved from [http://www.cdcr.ca.gov/adult\\_research\\_branch/Research\\_Documents/COMPAS\\_Final\\_Report\\_08-11-10.pdf](http://www.cdcr.ca.gov/adult_research_branch/Research_Documents/COMPAS_Final_Report_08-11-10.pdf)

### ***Federal Post Conviction Risk Assessment***

#### *Description*

The Federal Post Conviction Risk Assessment (PCRA) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations (Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011).

The PCRA contains 30 static and dynamic risk factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at work or school, and substance use problems. Self-report assessment scores are combined with probation officer assessment scores to arrive at an overall risk score. The PCRA is a 4<sup>th</sup> generation risk assessment instrument.

PCRA assessments comprise two components: 1) the Officer Assessment, and 2) Offender Self-Assessment. The self-report questionnaire consists of items that are “scored” and “unscored”. The 15 scored items are those that have been shown in studies conducted by the Administrative Office of U.S. Courts (Administrative Office) to predict recidivism and contribute to the overall risk score. The 15 unscored items have been shown in other research to predict recidivism, but have not been evaluated by the Administrative Office. They are included to inform intervention strategies, but do not contribute to the risk scores. Assessments must be administered by probation officers who have passed the online certification test created and offered by the Administrative Office; the Administrative Office prohibits uncertified assessors from accessing the PCRA. Prior to the online certification, probation officers must complete 16 hours of

training. They also must renew their certification every year. The PCRA is available through the Administrative Office at [www.uscourts.gov](http://www.uscourts.gov).

### *U.S. Research Evidence*

One study has assessed the predictive validity of PCRA assessments in two large U.S. samples. Johnson, Lowenkamp, VanBenschoten, and Robinson (2011) found excellent predictive validity in both. As of December 2012, there were no studies comparing predictive validity between assessments completed in research and practice contexts, by offender sex or by offender race/ethnicity. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

### *Practical Issues and Considerations*

Though promising, research evidence is limited to date. As noted above, there were no published evaluations of the reliability and predictive validity of PCRA assessments that met our inclusion criteria beyond the initial construction and validation study. However, a study published early this year by the instrument's authors (Lowenkamp, Johnson, VanBenschoten, & Robinson, 2013) compared predictive validity between research and practical contexts and reported high rates of inter-rater agreement. Independent replication is needed.

### *Selected References and Suggested Readings*

Administrative Office of the United States Courts, Office of Probation and Pretrial Services. (2011, September). *An overview of the Federal Post Conviction Risk Assessment*. Retrieved from [http://www.uscourts.gov/uscourts/FederalCourts/PPS/PCRA\\_Sep\\_2011.pdf](http://www.uscourts.gov/uscourts/FederalCourts/PPS/PCRA_Sep_2011.pdf)

Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The construction and validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, 75, 16-29.

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). *Psychological Services*, 10, 87-96.

## ***Inventory of Offender Risk, Needs, and Strengths***

### *Description*

The Inventory of Offender Risk, Needs, and Strengths (IORNS) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations (Miller, 2006a).

The IORNS contains 130 static, dynamic, risk, and protective factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, substance use problems, mental health problems,

and housing. Individual item responses are summed to create Static, Dynamic and Protective indexes as well as an Overall risk index. There also are two validity scales. The IORNS is a 3<sup>rd</sup> generation risk assessment instrument.

The IORNS is a true/false self-report questionnaire completed by the offender and requires 3<sup>rd</sup> grade reading level. The IORNS manual indicates that assessments take 15 to 20 minutes to administer, and 20 to 25 minutes to score. There are no training requirements for assessors, provided the purchaser of the exam has a degree in forensic or clinical psychology or psychiatry as well as certification in psychological testing. The purchaser also is responsible for overseeing the scoring of the assessment. IORNS assessments are available through Psychological Assessment Resources (parinc.com). Costs include those associated with the manual, interview guides, and assessment forms. For further information on pricing, see [www.parinc.com](http://www.parinc.com).

### *U.S. Research Evidence*

Predictive validity of IORNS assessments have been evaluated in only one U.S. sample conducted by the author of the instrument. Miller (2006b) found that offenders with higher Overall Risk Indices were in jail more frequently and had more non-violent arrests than those with lower scores. Similarly, those offenders who had more half-way house rule violations have significantly lower Overall Risk, and Dynamic Needs Indices.

As of December 2012, there were no published studies comparing predictive validity in U.S. samples between assessments completed in research and practice contexts, by recidivism outcome, offender sex, or offender race/ethnicity. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

### *Practical Issues and Considerations*

Though findings are promising, predictive validity of IORNS assessments has only been evaluated in one study conducted by the instrument developer that met our inclusion criteria; independent replication is needed.

### *Selected References and Suggested Readings*

Miller, H. A. (2006a). *Manual of the Inventory of Offender Risk, Needs, and Strengths (IORNS)*. Odessa, FL: Psychological Assessment Resources.

Miller, H. A. (2006b). A dynamic assessment of offender risk, needs, and strengths in a sample of pre-release general offenders. *Behavioral Sciences & the Law*, 24, 767-782.

## ***Level of Service Instruments***

### *Description*

The Level of Service family of instruments includes the Level of Service Inventory-Revised (LSI-R) and Level of Service Inventory-Revised: Screening Version (LSI-R:SV), actuarial risk assessment instruments intended to assess risk for general offending and violations across offender populations (Andrews & Bonta, 1995; 1998).

The LSI-R contains 54 static and dynamic risk factors. Content areas include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, leisure or recreational activities, substance use problems, mental health problems, and housing. Item responses are scored and summed for a total score from 0 to 54 that is used to classify risk as: Low = 0-23; Medium = 24-33; and High = >34. The LSI-R is a 3<sup>rd</sup> generation risk assessment instrument.

The LSI-R:SV contains eight static and dynamic items selected from the LSI-R. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, and substance abuse problems. Individual item responses are scored and summed for a total score ranging from 0-9. This score is used to determine if the offender requires a full LSI-R assessment. Like the interview-based version, the LSI-R:SV is also a 3<sup>rd</sup> generation risk assessment instrument.

LSI-R and LSI-R:SV assessments are completed through interview and file review, a process estimated to require approximately 30-40 minutes for the LSI-R and 10-15 minutes for the LSI-R:SV (though studies we reviewed reported longer completion times – see below). The assessor does not need formal training, but scoring must be overseen by someone who has post-secondary training in psychological assessment. The LSI-R and LSI-R:SV materials are available through Multi-Health Systems ([www.mhs.com](http://www.mhs.com)). Costs include those associated with the manual, interview guides, and assessment forms. For further information on pricing, see [www.mhs.com](http://www.mhs.com).

### *U.S. Research Evidence*

Predictive validity of LSI-R total scores had been evaluated in 25 U.S. samples as of December 2012. Performance in has ranged from poor to good, with the median on the cusp of fair and good. There were no studies examining the predictive validity of the risk classifications (as opposed to total scores) that met criteria for inclusion in this review. LSI-R total scores seem perform slightly better for men than for women, though performance is in the fair-good range for both. U.S. studies have not shown differences in validity as a function of racial/ethnicity. Predictive validity for total scores completed in the context of research and practice also is comparable. Validity in predicting is general offending is slightly better than violations. In the one U.S. study reporting inter-rater reliability data, agreement ranged from poor to excellent across content domains, but was excellent overall (Simourd, 2006).

Predictive validity of the LSI-R:SV has only been examined in two U.S. samples with mixed results: one study showed fair performance (Walters, 2011) and the other, good (Lowenkamp et



al., 2009). The LSI-R:SV seems to perform better for men (good predictive validity) than for women (fair predictive validity). There had been no studies comparing predictive validity between total scores and risk classifications, assessments completed in research and practice, by offender race/ethnicity, or by recidivism outcome as of December 2012. Because the LSI-R:SV is a self-report instrument, inter-reliability is not relevant.

The LSI-R instruments have been evaluated extensively outside of the United States. For example, there have been many evaluations of the predictive validity and inter-rater reliability of the LSI-R conducted in Canada and Europe (see, for example, Vose, Cullen, & Smith, 2008), but none have compared the predictive validity between total scores and risk classifications. Similarly, the LSI-R:SV has been studied outside of the United States (e.g., Daffern et al., 2005; Ferguson et al., 2005), but the research does not address the limitations described above.

### *Practical Issues and Considerations*

Researchers and professionals have reported administration times that deviate considerably from the LSI-R manual's estimate of 30-40 minutes, including 60 minutes in one sample (Holsinger et al., 2004) and 45-90 minutes in two others (Evans, 2009; Latessa et al., 2009).

There is considerable variation in the cut-off scores used for the risk categories. The manual encourages altering cut-off scores based on offense group characteristics, but research should be conducted *prior to* implementation to establish the validity of revised cut-off scores (Kim, 2010).

A recent addition to the Level of Service family of instruments is the Level of Service/Case Management Inventory (LS/CMI), an actuarial risk assessment with 43 items intended to aid professionals in offender management with late adolescent and adult offenders. No studies examining the LS/CMI met our inclusion criteria. However, there have been many evaluations of the predictive validity of the LS/CMI conducted outside of the United States (Andrews et al., 2011). Studies have included samples of male and female, as well as young offenders. Performance estimates for these populations ranged from fair to excellent. Inter-rater reliability has also been evaluated for total scores and found to be excellent (Rettinger & Andrews, 2010).

### *Selected References and Suggested Readings*

Andrews, D. A. & Bonta, J. (1995). *LSI-R: The Level of Service Inventory-Revised user's manual*. Toronto: Multi-Health Systems.

Andrews, D. A., & Bonta, J. L. (1998). *Level of Service Inventory-Revised: Screening Version (LSI-R:SV): User's manual*. Toronto: Multi-Health Systems.

Andrews, D. A., Bonta, J., Wormith, J. S., Guzzo, L., Brews, A., Rettinger, J., & Rowe, R. (2011). Sources of variability in estimates of predictive validity: A specification with Level of Service general risk and need. *Criminal Justice & Behavior*, 38, 413-432.

Daffern, M., Ogloff, J. R. P., Ferguson, M., & Thomson, L. (2005). Assessing risk for aggression in a forensic psychiatric hospital using the Level of Service Inventory-Revised: Screening Version. *International Journal of Forensic Mental Health, 4*, 201-206.

Ferguson, A. M., Ogloff, J. R. P., & Thomson, L. (2005). Predicting recidivism by mentally disordered offenders using the LSI-R:SV. *Criminal Justice & Behavior, 36*, 5-20.

Lowenkamp, C. T., Lovins, B., & Latessa, E. J. (2009). Validating the Level of Service Inventory-Revised and the Level of Service Inventory: Screening Version with a sample of probationers. *The Prison Journal, 89*, 192-204.

Rettinger, L. J., & Andrews, D. A. (2010). General risk and need, gender specificity, and the recidivism of female offenders. *Criminal Justice & Behavior, 37*, 29-46.

Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation, 72*, 22-29.

### ***Ohio Risk Assessment System***

#### *Description*

The Ohio Risk Assessment System (ORAS) is comprised of five actuarial risk assessment instruments intended to assess risk for recidivism across offender populations (Latessa et al., 2009): the 7-item Pretrial Assessment Tool (ORAS-PAT), the 4-item Community Supervision Screening Tool (ORAS-CSST), the 35-item Community Supervision Tool (ORAS-CST), the 31-item Prison Intake Tool (ORAS-PIT), and the 20-item Prison Re-entry Tool (ORAS-RT). Each includes static and dynamic risk factors and is designed for use at a specific stage in the criminal justice system; namely, pretrial, community supervision, institutional intake, and community reentry. Assessments identify criminogenic needs and place offenders into risk categories. An additional sixth instrument, the Prison Screening Tool (ORAS-PST), is designed to identify low risk inmates who do not need the full ORAS-PIT assessment.

Item responses are scored and summed to create total scores which are compared against risk classification cut-off values. The ORAS-PAT has a range from 0 to 9, the ORAS-CSST from 0 to 7, the ORAS-CST from 0 to 49, the ORAS-PIT from 3 to 29, and the ORAS-RT from 0 to 28. Each tool considers the offender's history of antisocial behavior, circumstance at school or work, and substance abuse problems; some also evaluate additional domains, such as attitudes (e.g., ORAS-CST, ORAS-RT), and mental health problems (e.g., ORAS-PIT, ORAS-RT). Together, the ORAS system reflects the 4<sup>th</sup> generation of risk assessment.

The ORAS tools are completed through a structured interview and analysis of official records; the ORAS-CSST, ORAS-PIT, and ORAS-RT additionally use self-report questionnaires. Assessors must complete a 2-day training package that accompanies the tool prior to administering any assessments. The ORAS is published by the Ohio Department of Rehabilitation and Correction (<http://www.drc.ohio.gov>). The system is non-proprietary and can

be obtained from the Center of Criminal Justice Research, University of Cincinnati (<http://www.uc.edu/corrections/services/risk-assessment.html>).

### *U.S. Research Evidence*

ORAS-PAT total scores demonstrated fair validity in predicting arrest in the construction sample and good validity in the validation sample (Latessa et al., 2009). A second evaluation found fair predictive validity for ORAS-PAT assessments, good validity for ORAS-PIT and ORAS-RT assessments, and excellent validity for ORAS-CCST and ORAS-CST assessments (Lowenkamp, Lemke, & Latessa, 2008). ORAS-PST assessments have not been included in these evaluations.

Predictive validity of ORAS assessments differs somewhat as a function of offender sex. Specifically, ORAS-CST assessments performed slightly better for male than female offenders, though predictive validity was excellent in both cases. Conversely, ORAS-PIT and ORAS-RT assessments performed better for female (excellent predictive validity) than male offenders (good). ORAS-CCST assessments, in contrast, have shown comparable predictive validity for both male and female offenders. The ORAS-PAT total scores have demonstrated better validity in predicting violations (good) than general offending (fair).

As of December 2012, there had been no U.S. studies comparing predictive validity between total scores and risk classifications, assessments completed in research and practice contexts, or by offender race/ethnicity that met our inclusion criteria. There also had not been any evaluations of inter-rater reliability.

### *Practical Issues and Considerations*

Though findings are very promising, there has been relatively little research on the predictive validity of the ORAS, with only one evaluation of four of the tools and two of the other. Further, studies that met our inclusion criteria did not report inter-rater reliability of the assessments. Finally, all research on the ORAS reviewed in this report had been completed by the study developers; independent replication is needed.

### *Selected References and Suggested Readings*

Latessa, E., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). *Creation and validation of the Ohio Risk Assessment System: Final report*. Cincinnati, OH: Authors. Retrieved from [http://www.uc.edu/ccjr/Reports/ProjectReports/ORAS\\_Final\\_Report.pdf](http://www.uc.edu/ccjr/Reports/ProjectReports/ORAS_Final_Report.pdf)

Lowenkamp, C. T., Lemke, R., & Latessa, E. (2008). The development and validation of a pretrial screening tool. *Federal Probation*, 72, 2-9.

## ***Risk Management Systems***

### *Description*

The Risk Management Systems (RMS) is an actuarial risk assessment instrument intended for use intended to assess risk for general offending across offender populations (Dow, Jones, & Mott, 2005). The RMS currently contains 67 static and dynamic risk factors; however, when it was validated, the instrument included only 65 items. The assessment is split into four parts: 1) Needs (24 items), 2) Risk (9 items), 3) Mental Health (10 items), and 4) Other-External (24 items). Content areas assessed include attitudes, associates or peers, history of antisocial behavior, personality problems, relationships, circumstances at school or work, substance abuse problems, mental health problems, and housing. The developers of the RMS describe it as a 5<sup>th</sup> generation risk assessment instrument due to its exemplar-based approach.

The RMS is administered using a computer-based questionnaire. As such, the assessor is removed from the initial assessment process; individual item responses are statistically analyzed to calculate risk of recidivism. Risk scores for violence and recidivism range from 1.00 (Low) to 2.00 (High), at 0.01 intervals. However, there are no established cut-off scores for risk categories, so the assessor must interpret the subsequent level of risk/supervision required. RMS assessment materials are available through Syscon Justice Systems ([www.syscon.net](http://www.syscon.net)). For information on pricing see [www.syscon.net](http://www.syscon.net).

### *U.S. Research Evidence*

As of December 2012, predictive validity of RMS assessments had been reported in two U.S. studies; performance ranged from good (Kelly, 2009; later republished in Shaffer et al., 2010) to excellent (Dow et al., 2005). The risk classifications have notably better predictive validity (excellent) compared to total scores (good). Validity is comparable for predicting general offending and violations. RMS assessments appear to have better predictive validity when completed in research studies (excellent) than in the context of 'real world' practice (good); however, risk classifications were used in one study and total scores in the other.

There were no studies of predictive validity conducted in the United States that compared findings across offender sex or racial/ethnic groups. There also were no U.S. evaluations of inter-rater reliability that met our inclusion criteria.

### *Practical Issues and Considerations*

In the initial development and validation work, the tool was intended to be used for assessing risk for general offending (Dow et al., 2005), but a later study established the validity of RMS assessments in predicting violations (Kelly, 2009). Overall, further independent research is needed to replicate and establish the generalizability of findings, as well as to determine the validity of different cut-off scores.

### *Selected References and Suggested Readings*

Dow, E., Jones, C., & Mott, J. (2005). An empirical modeling approach to recidivism classification. *Criminal Justice and Behavior*, 32, 223-247.

Kelly, B. (2009). *A validation study of Risk Management Systems* (Master's thesis). Retrieved from UNLV Theses/Dissertations/Professional Papers/Capstones. (Paper 128). <http://digitalscholarship.unlv.edu/thesesdissertations/128>

Shaffer, D. K., Kelly, B., & Lieberman, J. D. (2010). An exemplar-based approach to risk assessment: Validating the Risk Management Systems instrument. *Criminal Justice Policy Review*, 22, 167-186.

### ***Salient Factor Score***

#### *Description*

The Salient Factor Score (SFS) is an actuarial risk assessment tool intended to inform decisions regarding whether an offender should be granted parole or not. The SFS is a 2<sup>nd</sup> generation risk assessment instrument.

There are at least four versions of the SFS, all of which measure static risk factors. Items have been adapted throughout the years to be consistent with research findings. The SFS74 contains nine items and content areas include history of antisocial behavior, circumstances at work or school, substance use problems, and housing. The SFS76 contains seven items and content areas include history of antisocial behavior, circumstances at work or school, and substance use problems. The SFS81 contains six items and content areas include history of antisocial behavior and substance use problems. The SFS98 includes six items and the only content area included is history of antisocial behavior. Unlike the prior versions, the SFS98 also considers whether the offender was older than 41 at the time of the current offense.

SFS assessments are completed through review of official records. Item ratings are summed to arrive at an overall risk score; a *higher* score indicating *lower* risk. These total scores are then used to place offenders within one of four risk categories: very good risk, good risk, fair risk, and poor risk. For further information contact the United States Parole Commission (<http://www.justice.gov/uspc>).

#### *U.S. Research Evidence*

As of December 2012, predictive validity of SFS74, SFS76, and the SFS81 assessments had been examined in 15 U.S. samples. Validity of SFS74 and SFS76 assessments in predicting general offending has ranged from good to excellent. SFS81 assessments also have shown excellent predictive validity across most studies, though the odds ratio was notably low in one evaluation (Howard, 2007). We did not find any evaluations of the predictive validity of SFS98 assessments that met our inclusion criteria.

To date, there have been no U.S. studies comparing predictive validity of the SFS instruments between total scores and risk classifications, assessments completed in research and practice contexts, or by offender race/ethnicity. We also did not find any evaluations of inter-rater reliability that met our inclusion criteria.

### *Practical Issues and Considerations*

Though items are relatively straightforward to code, investigations of inter-rater reliability are needed to establish the consistency of assessments completed by different assessors.

Jurisdiction-specific adaptations include the Connecticut Salient Factor Score.

### *Selected References and Suggested Readings*

Hoffman, P. (1996). Twenty years of operational use of a risk prediction instrument: The United States Parole Commission's Salient Factor Score. *Journal of Criminal Justice*, 22, 477-494.

Hoffman, P. & Adelberg, S. (1980). The Salient Factor Score: A nontechnical overview. *Federal Probation*, 44, 44-52.

Howard, B. (2007). *Examining predictive validity of the Salient Factor Score and HCR-20 among behavior health court clientele: Comparing static and dynamic variables*. (Unpublished doctoral dissertation).

### ***Self-Appraisal Questionnaire***

The Self-Appraisal Questionnaire (SAQ) is an actuarial risk assessment instrument to assess risk for general offending among male offenders (Loza, 2005).

The SAQ contains 72 dynamic and static risk factors. Content areas include attitudes, associates or peers, history of antisocial behavior, personality problems, and substance abuse problems. Items are divided across seven subscales. Scores on six subscales are calculated to provide an overall risk score. A seventh anger subscale is not used to assess risk for recidivism. Therefore, of the 72 total items, 67 items are used to predict recidivism. Total scores are used to place offenders in one of four risk categories: low, low-moderate, high-moderate, and high. The SAQ is a 3<sup>rd</sup> generation risk assessment instrument.

The SAQ is a true/false self-report questionnaire. Five items can be used to assess the validity of an offender's answers by comparing them against official records. The SAQ takes approximately 15 minutes to administer and five minutes to hand-score. The assessor does not need formal training, but scoring must be overseen by someone who has post-secondary training in psychological assessment. The SAQ can be purchased from Multi-Health Systems Inc. at [www.mhs.com](http://www.mhs.com). Costs include those associated with the manual and assessment forms. For further information on pricing, see [www.mhs.com](http://www.mhs.com).

### *U.S. Research Evidence*

Two studies have evaluated the predictive validity of the SAQ in U.S. samples. These studies used low, moderate, and high risk categories rather than the four categories suggested by the assessment developer. Mitchell and Mackenzie (2006) found poor validity of the SAQ

assessments in predicting re-arrest and failed to find differences in total scores between recidivists and non-recidivists. In contrast, using a longer follow-up period and a larger sample, Mitchell, Caudy and Mackenzie (2012) found that SAQ assessments predicted time to first reconviction, though the effect size was small.

As of December 2012, there had been no studies comparing predictive validity in U.S. samples between total scores and risk classifications, assessments completed in research and practice, by offender sex, or race/ethnicity that met our inclusion criteria. Because the SAQ is a self-report instrument, inter-reliability is not relevant.

There have been many evaluations of the SAQ in Canada (e.g., Kroner & Loza, 2001; Loza & Loza-Fanous, 2000; Loza et al., 2005), but none have compared the predictive validity between total scores and risk classifications, research and practice contexts, by offender sex, or race/ethnicity.

### *Practical Issues and Considerations*

The SAQ requires a 5<sup>th</sup> grade reading level. Prior studies of the validity of SAQ assessments in predicting violent outcomes, including institutional violence and violent recidivism (e.g., Campbell, French & Gendreau, 2009), as well as violent and non-violent recidivism in Canadian samples (e.g., Loza, MacTavish, & Loza-Fanous, 2007) have shown more promising results than those reported herein vis-à-vis validity in predicting non-violent offending in U.S. samples.

### *Selected References and Suggested Readings*

Kroner, D., & Loza, W. (2001). Evidence for the efficacy of self-report in predicting violent and nonviolent criminal recidivism. *Journal of Interpersonal Violence, 16*, 168-177.

Loza, W., & Loza-Fanous, A. (2000). Predictive validity of the Self-Appraisal Questionnaire (SAQ): A tool for assessing violent and nonviolent release failures. *Journal of Interpersonal Violence, 15*, 1183-1191.

Loza, W. (2005). *The Self-Appraisal Questionnaire (SAQ): A tool for assessing violent and non-violent recidivism*. Toronto: Mental Health Systems.

Loza, W., Neo, L. H., Shahinfar, A., & Loza-Fanous, A. (2005). Cross-validation of the Self-Appraisal Questionnaire: A tool for assessing violent and nonviolent recidivism with female offenders. *International Journal of Offender Therapy & Comparative Criminology, 49*, 547-560.

Mitchell, O., Caudy, M., & Mackenzie, D. (2012). A reanalysis of the Self-Appraisal Questionnaire: Psychometric properties and predictive validity. *International Journal of Offender Therapy and Comparative Criminology 20*, 1-15.

Mitchell, O., & Mackenzie, D. (2006). Disconfirmation of the predictive validity of the Self-Appraisal Questionnaire in a sample of high-risk drug offenders. *Criminal Justice and Behavior 33*, 449-466.

## *Service Planning Instruments*

### *Description*

The Service Planning Instrument (SPIn) is an actuarial risk assessment tool intended to assess risk for offending and to identify service needs of male offenders. The SPIn-W was developed for use with female offenders.

Both the SPIn and SPIn-W are self-report, computer-based instruments. The SPIn includes 90 static, dynamic, risk, and protective factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at school or work, substance use problems, mental health problems, and housing. The SPIn-W includes 100 static, dynamic, risk, and protective factors. Content areas include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at school or work, leisure or recreational activities, substance use problems, mental health problems, and housing. The SPIn and SPIn-W are 4<sup>th</sup> generation risk assessment instruments.

For both instruments, software is used to calculate an offender's risk score which is presented graphically and narratively. The assessor must compare responses on static items to the offender's official records. Assessors are required to attend a two-day training session. Additional 2-day training program to help administrators better prepare for the case planning process, as well as data workshops, refresher courses, technical support, and quality assurance also are available. The SPIn and SPIn-W can be purchased from Orbis Partners Inc. ([www.orbispartners.com](http://www.orbispartners.com)). For information on pricing, see [www.orbispartners.com](http://www.orbispartners.com).

### *U.S. Research Evidence*

As of December 2012, there were no published studies assessing predictive validity of SPIn assessments in U.S. samples. Two studies have evaluated predictive validity of the SPIn-W assessments; performance ranged from poor to excellent.

There were no comparisons of predictive validity in U.S. samples between total scores and risk classifications, assessments completed in research and practice contexts, by outcome or by offender race/ethnicity that met our inclusion criteria. We also did not identify any U.S. evaluations of inter-rater reliability that met these criteria.

### *Practical Issues and Considerations*

Current evidence regarding the predictive validity of SPIn-W assessments is both limited and mixed. More research is needed.

### *Selected References and Suggested Readings*

Meaden, C. (2012). *The utility of the Level of Service Inventory-Revised versus the Service Planning Instrument for Women in predicting program completion in female offenders.*



(Unpublished Master's thesis). Retrieved from Central Connecticut State University Theses, Dissertations, and Special Projects.

Millson, B., Robinson, D., & Van Dietsen, M. (2010). *Women Offender Case Management Model: An outcome evaluation*. Washington, DC: U.S. Department of Justice, National Institute of Corrections. Retrieved from:

<http://www.cjinvolvedwomen.org/sites/all/documents/Women%20Offender%20Case%20Management%20Model.pdf>

### ***Static Risk and Offender Needs Guide***

The Static Risk and Offender Needs Guide (STRONG) is an actuarial risk assessment instrument intended to assess risk for general offending across offender populations (Barnoski & Drake, 2007).

The STRONG consists of three parts: 1) the Static Risk Assessment which contains 26 static risk factors; 2) the Offender Needs Assessment which contains 70 dynamic risk and protective factors; and 3) the Offender Supervision Plan, which is auto-populated based on the results of the Offender Needs Assessment. Content areas assessed in the Static Risk Assessment include history of antisocial behavior and substance use problems. Items scores are used to create three separate scores: Felony Risk Score; Non-Violent Felony Risk Score (high property risk/high drug risk); and Violent Felony Risk Score. These three scores are used to classify offenders in one of five categories: high risk violent; high risk property; high risk drug; moderate risk; and low risk. Content areas assessed in the Offender Needs Assessment include attitudes, associates or peers, personality problems, relationships, circumstances at work or school, substance use problems, mental health problems, and housing. Ratings on items included in the Offender Needs Assessment are not used to inform risk assessments, but instead guide the development of interventions designed to reduce risk of future criminal justice involvement. As such, the STRONG is a 4<sup>th</sup> generation risk assessment instrument.

STRONG assessments are completed by assessors using a web-based interface. Assessors must complete an initial training program as well as routine booster training sessions. The STRONG was developed by Assessments.com in collaboration with the Washington Department of Corrections. A very similar version can be purchased for use in other jurisdictions through [www.assessments.com](http://www.assessments.com).

### ***U.S. Research Evidence***

Only one study that met our inclusion criteria has evaluated the predictive validity of STRONG assessments; assessments demonstrated excellent predictive validity overall as well as for male and female offenders separately (Barnoski & Drake, 2007). There were no U.S. studies comparing predictive validity as a function of offender race/ethnicity, type of recidivism outcome or between assessments completed in the context of research versus practice. We also did not find any evaluations of inter-rater reliability that met inclusion criteria.

### *Practical Issues and Considerations*

Though findings are promising, predictive validity of STRONG assessments has only been evaluated in one study conducted by the instrument developer; independent replication is needed.

### *Selected References and Suggested Readings*

Barnoski, R., & Drake, E. K. (2007). *Washington's Offender Accountability Act: Department of Corrections' static risk instrument*. Olympia, WA: Washington State Institute for Public Policy. Retrieved from <http://www.wsipp.wa.gov/rptfiles/07-03-1201R.pdf>

### ***Wisconsin Risk and Needs Scales***

#### *Description*

The Wisconsin Risk and Needs scales (WRN) is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations. A revised version (WRN-R) was designed specifically for use with probationers and parolees (Eisenberg, Bryl, & Fabelo, 2009). Both the WRN and WRN-R are 4<sup>th</sup> generation risk assessment instruments.

The WRN contains 53 static and dynamic risk factors. Content areas assessed include attitudes, associates or peers, history of antisocial behavior, relationships, circumstances at work or school, substance use problems, and mental health problems. Individual item scores are scored and summed for a total risk score ranging from 0 to 52. The total score is used to place the offender in a risk category based on predetermined cut-offs: Low = 0-7; Medium = 8-14; and High = 15+.

The WRN-R retained 52 of the WRN's items and covers the same content areas. The weights of the different factors have been revised from the original WRN based on the results of a validation study, and the revised total risk score has a range of 0 to 25. The total score is used to estimate risk level based on new cut-offs: Low = 0-8; Medium = 9-14; and High = 15+.

WRN assessments are completed using information obtained through interview. The WRN is non-proprietary and available through Justice Systems Assessment & Training (<http://www.jsatresources.com/Toolkit/Adult/adf6e846-f4dc-4b1e-b7b1-2ff28551ce85>).

#### *U.S. Research Evidence*

Predictive validity of the WRN assessments have ranged from fair (Eisenberg et al., 2009) to excellent (Connolly, 2003). WRN assessments appear to perform better when predictive violations (excellent) than general offending (good). Our comparisons between predictive validity of assessments completed in research versus practice failed to identify any differences. As of December 2012, no U.S. studies compared predictive validity between WRN total scores and risk classifications, by offender sex, or race/ethnicity. We also did not identify any U.S. evaluations of inter-rater reliability that met our inclusion criteria.

As of December 2012, predictive validity of WRN-R assessments had been evaluated in one U.S. study; assessments demonstrated good predictive validity. To date, there have been no studies comparing predictive validity in U.S. samples between WRN-R total scores and risk classifications, assessments completed in research and practice contexts, by recidivism outcome, offender race/ethnicity, or sex that met our inclusion criteria. We also did not identify any U.S. evaluations of inter-rater reliability of WRN-R assessments.

### *Practical Issues and Considerations*

A high percentage of offenders are classified as high risk using the WRN due to the heavy weight given to convictions for an assaultive offense in the past five years. There is concern that such over-classification is “counter to the goal of risk classification: to differentiate the population by risk and allocate resources accordingly” (Eisenberg et al., 2009, p. iv).

In 2004, a new, automated assessment and case management system called the Correctional Assessment and Intervention System (CAIS) was developed based upon the WRN and the Client Management Classification tools (Baird, Heinz, & Bemus, 1979). This CAIS is an actuarial risk assessment instrument intended to assess risk for general offending and violations across offender populations, as well as to be used in the development of case management plans. Its predictive validity has not yet been evaluated.

### *Selected References and Suggested Readings*

Baird, C., Heinz, R., & Bemus, B. (1979). *The Wisconsin Case Classification/Staff Deployment Project*. Madison, WI: Wisconsin Department of Corrections.

Eisenberg, M., Bryl, J., & Fabelo, T. (2009). *Validation of the Wisconsin Department of Corrections risk assessment instrument*. New York: Council of State Governments Justice Center. Retrieved from [http://www.wi-doc.com/PDF\\_Files/WIRiskValidation\\_August%202009.pdf](http://www.wi-doc.com/PDF_Files/WIRiskValidation_August%202009.pdf)

## OTHER TYPES OF INSTRUMENTS USED TO ASSESS RECIDIVISM RISK

### *Violence Risk Assessment Instruments*

Violence risk assessment instruments, such as the Historical-Clinical-Risk Management-20 (HCR-20; Webster, Douglas, Eaves, & Hart, 1997) and Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 2006), are intended to assess risk of future violence specifically, but also are frequently used to assess risk of (non-violent) recidivism.

#### *HCR-20*

The HCR-20 is a structured professional judgment scheme comprised of 20 static and dynamic items that assess historical risk factors, clinical risk factors, and risk management factors. The individual item ratings are used to inform a final professional judgment of low, moderate, or high risk. Only one study has evaluated the validity of HCR-20 assessments in predicting recidivism in a U.S. sample (Barber-Rioja, Dewey, Kopelovich, & Kucharski, 2012). Overall, the assessment total score was found to have excellent validity in predicting both general offending and violations. The HCR-20 has been widely validated outside of the U.S. (see <http://kdouglas.files.wordpress.com/2007/10/hcr-20-annotated-biblio-sept-2010.pdf>).

#### *VRAG*

The VRAG is an actuarial instrument designed for use with previously violent, mentally disordered offenders. It consists of 12 items that gather information on static and dynamic risk factors. Individual item responses are weighted and summed for a total score, which is then used to estimate level of risk based on an actuarial table. The predictive validity of VRAG assessments for both general offending and violations also has been evaluated in only one U.S. sample (Hastings et al., 2011). Validity in predicting general offending ranged from good to excellent for male offenders, and fair to good for female offenders. Validity in predicting violations ranged from fair to good for male offender and poor to fair for female offenders. Like the HCR-20, much research completed outside of the U.S. has examined the validity of VRAG assessments. For more information, visit <http://www.mhcop.on.ca/>

### *References and Suggested Readings*

Barber-Rioja, V., Dewey, L., Kopelovich, S., & Kucharski, L. T. (2012). The utility of the HCR-20 and PCL:SV in the prediction of diversion noncompliance and reincarceration in diversion programs. *Criminal Justice and Behavior*, *39*, 475-492.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). The prediction of violence and antisocial behaviour: A systematic review and meta-analysis of the utility of risk assessment instruments in 73 samples involving 24,827 individuals. *British Medical Journal*, *345*, e4692.

Hastings, M. E., Krishnan, S., Tangney, J. P., & Stuewig, J. (2011). Predictive and incremental validity of the Violence Risk Appraisal Guide scores with male and female jail inmates. *Psychological Assessment, 23*, 174-183.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). *HCR-20: Assessing risk for violence* (version 2). Burnaby, BC: Simon Fraser University, Mental Health, Law, and Policy Institute.

### ***Personality Assessment Instruments***

Personality assessment instruments, such as the Psychopathy Checklist-Revised (PCL-R; Hare, 2003), the Psychopathy Checklist: Screening Version (PCL:SV; Hart, Cox, & Hare, 1995), and the Personality Assessment Instrument (PAI; Morey, 1991), evaluate personality constructs that correlate with criminal offending (for a meta-analytic review see Singh & Fazel, 2010).

#### *PCL Instruments*

The PCL-R is a 20-item actuarial assessment that can be used to diagnosis psychopathy, a form of antisocial personality disorder characterized by a persistent pattern of severe and refractory callous-unemotionality. Individual items are scored through file review and semi-structured interview, then summed for total score ranging from 0 to 40 (where 30+ indicates the presence of psychopathy). The PCL:SV is a shorter, 12-item version. Again, individual item ratings are scored and summed, with a cutoff score of 18 typically used for classification of psychopathy. Research demonstrates excellent correspondence between the two measures in correctional samples (Guy & Douglas, 2006). Validity of PCL-R and PCL:SV assessments in predicting recidivism has been evaluated extensively in the U.S., with performance ranging from poor to good (e.g., Gonsalves, Scalora, & Huss, 2009; Salekin, Rogers, Ustad, & Sewell, 1998; Walters & Duncan, 2005). For more information on the PLC-R and PCL:SV, see <http://www.hare.org/scales/>.

#### *PAI*

The PAI contains 344 self-report items that are divided into 22 validity, clinical, treatment consideration, and interpersonal scales. Individual item responses within the scales are hand scored and assessed in conjunction with interpretive guidelines included in the professional manual (Morey, 2007). In U.S. studies assessing the predictive validity of the PAI, the assessment scale scores had fair to good validity in predicting general offending (e.g., Barber-Rioja et al., 2012; Walters, 2009; Walters & Duncan, 2005). For an overview and bibliography, see <http://www4.parinc.com/Products/Product.aspx?ProductID=PAI>.

### *Other Personality Assessment Instruments*

Other instruments including the California Psychological Inventory: Socialization Scale (CPI:SO), Lifestyle Criminality Screening Form (LCSF), Minnesota Multiphasic Personality Inventory (MMPI), Neuroticism, Openness to Exposure Personality Inventory-Revised (NEO-PI-R), and the Peterson, Quay, and Cameron Psychopathy Scale (PQC) can produce valid assessments of recidivism risk, though performance varies widely (see Walters, 2003, 2006).

### *References and Suggested Readings*

Barber-Rioja, V., Dewey, L., Kopelovich, S., & Kucharski, L. T. (2012). The utility of the HCR-20 and PCL:SV in the prediction of diversion noncompliance and reincarceration in diversion programs. *Criminal Justice and Behavior, 39*, 475-492.

Hare, R. D. (2003). *Hare Psychopathy Checklist-Revised (PCL-R): Second edition, technical manual*. Toronto, ON, Canada: Multi-Health Systems.

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version* (1st ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Gonsalves, V. M., Scalora, M. J., & Huss, M. T. (2009). Prediction of recidivism using the Psychopathy Checklist-Revised and the Psychological Inventory of Criminal Thinking Styles within a forensic sample. *Criminal Justice and Behavior, 36*, 741-756.

Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Salekin, R. T., Rogers, R., Ustad, K. L., & Sewell, K. W. (1998). Psychopathy and recidivism among female inmates. *Law and Human Behavior, 22*, 109-128.

Walters, G. D. (2009). The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening Version as incrementally valid predictors of recidivism. *Law and Human Behavior, 33*, 497-505.

Walters, G. D. & Duncan, S. A. (2005). Use of the PCL-R and PAI to predict release outcome in inmates undergoing forensic evaluation. *Journal of Forensic Psychiatry and Psychology, 16*, 459-476.

### *Criminal Thinking Questionnaires*

Criminal thinking questionnaires, such as the Psychological Inventory of Criminal Thinking Styles (PICTS; Walters, 1995) and the Texas Christian University Criminal Thinking Scales (TCU CTS; Knight, Simpson, & Morey, 2002), are designed to identify attitudes and thought patterns associated with criminal behavior.

#### *PICTS*

The PICTS is an 80-item, self-report measure composed of eight thinking pattern scales, two validity scales, four factor scales, two composite scales, and a General Criminal Thinking (GCT) scale. The validity of PICTS scores in predicting general offending has been evaluated in a number of U.S. studies with mixed findings. Performance of the GCT scale scores ranges from poor to good (e.g., Walters, 2009a, 2009b, 2011); however, other research suggests the eight thinking pattern scales have poor validity (Gonsalves, Scalora, & Huss, 2009).

#### *TCU CTS*

The TCU CTS is an actuarial, self-report instrument designed to measure criminal thinking. The instrument contains 37 items distributed across six thinking pattern scales: Entitlement, Justification, Power Orientation, Cold Heartedness, Criminal Rationalization, and Personal Irresponsibility. In one U.S. study, the six thinking pattern scale scores had poor validity in predicting both general offending and violations (Taxman, Rhodes & Dumenci, 2011). More information and a copy of the TCU CTS assessment materials are available from <http://www.ibr.tcu.edu/pubs/datacoll/cjtrt.html>.

#### *References and Suggested Readings*

Gonsalves, V. M., Scalora, M. J., & Huss, M. T. (2009). Prediction of recidivism using the Psychopathy Checklist-Revised and the Psychological Inventory of Criminal Thinking Styles within a forensic sample. *Criminal Justice and Behavior*, 36, 741-756.

Knight, K., Simpson, D. D., & Morey, J. T. (2002). *TCU-NIC Cooperative Agreement: Final report*. Fort Worth, TX: Texas Christian University, Institute of Behavioral Research.

Taxman, F. S., Rhodes, A. G., & Dumenci, L. (2011). Construct and predictive validity of Criminal Thinking Scales. *Criminal Justice and Behavior*, 38, 174-187.

Walters, G. D. (1995). The Psychological Inventory of Criminal Thinking Styles, Part I: Reliability and preliminary validity. *Criminal Justice and Behavior*, 22, 307-325.

Walters, G. D. (2009a). Effect of a longer versus shorter test-release interval on recidivism prediction with the Psychological Inventory of Criminal Thinking Styles (PICTS). *International Journal of Offender Therapy and Comparative Criminology*, 53, 665-678.

Walters, G. D. (2009b). The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening Version as incrementally valid predictors of recidivism. *Law and Human Behavior*, 33, 497-505.

Walters, G. D. (2011). Predicting recidivism with the Psychological Inventory of Criminal Thinking Styles and Level of Service Inventory-Revised: Screening Version. *Law and Human Behavior*, 35, 211-220.



## CONCLUSION

### *Summary of Findings*

Our review of validation studies conducted in the United States did not identify one instrument that systematically produced more accurate assessments than the others. However, performance within and between instruments varied considerably depending on the assessment sample, circumstances, and recidivism outcome.

Overall, there were very few U.S. evaluations examining the predictive validity of assessments completed using instruments commonly used in U.S. correctional agencies. In most cases, validity of assessments completed using any given instrument had only been examined in one or two studies conducted in the United States, and frequently, those investigations were completed by the same people who developed the instrument. Moreover, only two of the 53 studies included in this review reported evaluations of inter-rater reliability. (We return to these two points later.)

Our selection criteria and, specifically, our focus on studies of predictive validity conducted in the United States resulted in the exclusion of some prominent and promising instruments, such as the LS/CMI or the Women's Risk/Need Assessment. Similarly, none of the reviewed studies examined the predictive validity of structured professional judgment, as opposed to actuarial, instruments, though we know of at least a few that are being used for the purposes of assessing recidivism risk (e.g., the Short-Term Assessment of Risk and Treatability, START, see Desmarais, Van Dorn, Telford, Petrila, & Coffey, 2012). Importantly, findings of the current review are not intended to suggest that these instruments do not produce reliable and valid assessments of recidivism risk and should not necessarily preclude their use in practice. Instead, we are simply asserting that they have yet to be evaluated as such in the United States. Indeed, decision makers interested in any risk assessment instrument should balance considerations of the empirical evidence, but also the practical issues we review in the following section.

Finally, risk classifications (e.g., identification of offenders as low, moderate, or high risk) generally outperformed total scores, yet total scores were evaluated much more frequently. This finding is consistent with prior research (e.g., Desmarais et al., 2012) and emphasizes the importance of using the instruments as they were designed to be used.

### *Selecting a Recidivism Risk Assessment Instrument*

When deciding which recidivism risk assessment instrument to implement in practice, we recommend reviewing the empirical evidence, as well as answering the following questions:

*What is your outcome of interest?*

Our review revealed that some instruments performed better in predicting particular recidivism outcomes than others. Specifically, the SFS instruments performed particularly well in predicting general offending *including* violations, whereas the ORAS-CST, ORAS-CSST, PCRA, and

STRONG were excellent predictors of offenses *excluding* violations. WRN assessments stood out as the best predictors of violations alone.

*What is your population?*

Some instruments were developed to assess for specific populations; for example, the SFS instruments are specifically designed for use with parolees. Also, some instruments appear to perform better for some subgroups of offenders than others. The LSI instruments, for instance, produced assessments with only fair validity for female offenders, though predictive validity was generally good for male offenders. Other instruments, such as the COMPAS, ORAS and STRONG, produced assessments with good validity for both male and female offenders.

*What resources are required to complete the assessment?*

Answering this question includes considering characteristics of both the risk assessment tool as well as the setting; for instance, the information necessary to complete the assessment and whether this information is available. Some instruments, such as the IORNS, are completed based solely on offender self-report; other instruments, such as the PCRA and COMPAS, combine information derived from a variety of sources, including self-report, interview, and review of official records. Similarly, the time required to complete a risk assessment will depend not only on the nature and amount of information required, but also the number of items included. We found that the number of items varied broadly across instruments from four items (ORAS-CSST) to 130 items (IORNS). Decision makers should consider whether staff have the time and information required to complete the assessments. Other resource considerations include staff training and backgrounds. Some instruments, such as the PCRA, require that assessors complete training courses and are certified prior to implementation. Others, such as the LSI family of instruments, require that assessors be supervised by professionals with specific degrees and/or credentials. Last, but certainly not least, decision makers should consider the costs associated with implementing any given risk assessment tool. Costs may include those associated with purchasing materials and staff training, among others, and they may be fixed, one-time costs or costs that will continue to be incurred over time. Long-term sustainability of implementation will hinge, in part, on a *realistic* appraisal of the match between the available and required resources.

***Additional Considerations***

In addition to identifying the instrument best-suited to an agency's specific needs and constraints, there are additional issues to consider during the process of selecting and implementing a recidivism risk assessment tool.

First, caution is warranted when attempting to generalize the findings of research studies to the use of risk assessment instruments in practice. In research contexts, risk assessments are routinely conducted by graduate students, who may have more or less training than those who will be conducting the risk assessments in practice. Assessors in research studies also may be given more time and resources to complete risk assessments and may receive ongoing

supervision in the specific risk assessment protocol; these luxuries typically are not afforded to professionals in practice settings.

Second, there have been very few evaluations of predictive validity within specific offender subgroups. Indeed, only a handful of studies included in this review compared validity depending on offender sex or race/ethnicity and none examined predictive validity across psychiatric diagnostic categories. As such, there is insufficient evidence to conclude that assessments perform comparably or are equally applicable to specific offender subgroups. As described earlier, actuarial instruments estimate risk of recidivism through comparison of a given offender's total score against the recidivism rates of offenders with the same (or a similar) score in the construction sample. Race/ethnicity and sex are important factors associated with recidivism that may not be accounted for in these actuarial models. There is considerable evidence to suggest that race/ethnicity and sex are potentially important sources of assessment bias (Holtfreter & Cupp, 2007; Leistico, Salekin, DeCoster, & Rogers, 2008).

Third, allegiance, which occurs when at least one developer of the risk assessment instrument is an author on a study investigating that instrument's predictive validity, was present for many of the articles included in this review. Strong effects of allegiance on evaluations of assessment and treatment approaches, including risk assessment, have been found in many fields. In the violence risk assessment literature, a recent meta-analysis demonstrated the impact of allegiance on the predictive validity of three commonly used actuarial instruments (Blair, Marcus, & Boccaccini, 2008). Performance of the instruments was significantly better in studies conducted by the tool authors than in studies conducted by independent researchers. We were unable to test for allegiance effects due to the relatively small number of studies per instrument. Though the reasons for allegiance effects are unclear (e.g., bias, fidelity, see Harris & Rice, 2010), there is a critical need for independent evaluation of the predictive validity of risk assessments completed using the instruments included in this review.

Fourth, most studies included in this review reported statistics that speak to whether recidivists generally received higher risk estimates than did non-recidivists (known as *discrimination*). Very few studies reported statistics that speak to whether those offenders who were identified as high risk for recidivism went on to recidivate during follow-up and whether those offenders who were identified as low risk did not (known as *calibration*). This is not unique to the studies included in the current review; a recent review found that calibration estimates were reported in less a fourth of violence risk assessment studies (see Singh, Desmarais & Van Dorn, 2013). Discrimination and calibration are two sides of the same coin – both representing important qualities of an instrument's predictive validity – but address different issues (Singh, 2013).

Fifth, there was an almost complete lack of information regarding the inter-rater reliability of available recidivism risk assessment instruments. With the exception of LSI-R and LSI-R:SV, we do not have any information regarding whether assessments completed using the instruments reviewed in this report are consistent across assessors. This is not trivial; reliability has been referred to as “the most basic requirement for a risk assessment instrument” (Douglas, Nicholson, & Skeem, 2011, p. 333). Indeed, an assessment *must* be reliable in order for it to be valid (though the reverse is not true). Inter-rater reliability is relevant to any assessment in which

an assessor must rate or code items as part of the process; thus, inter-rater reliability should be examined for all instruments except those completed exclusively through offender self-report.

Sixth and finally, there have been few evaluations of the impact of implementing a risk assessment tool on recidivism rates. Though many of the instruments included in the present review have acceptable levels of predictive validity, the goal of risk assessment is not simply to predict, but, ultimately, to *reduce* recidivism. Achieving this goal will necessitate the following:

1. The risk assessment tool *must* be implemented in a sustainable fashion with fidelity. It is not as simple as deciding on a tool and applying it in practice. Successful implementation of a risk assessment tool involves completing a series of steps, from preparation to training and pilot testing to full implementation. This multi-step process requires ongoing supervision to ensure sustainability, including regular evaluations of fidelity and booster training for staff on a semi-annual basis (see Vincent, Guy & Grisso, 2012 for a guide to implementation).
2. Findings of the risk assessment *must* be communicated accurately and completely. Indeed, “Improper risk communication can render a risk assessment that was otherwise well-conducted completely useless or even worse, if it gives consumers the wrong impression.” (Heilbrun, Dvoskin, Hart & McNiel, 1999, p. 94).
3. Information derived during the risk assessment process *must* be used to guide risk management and rehabilitation efforts, with particular attention to the steps described by the RNR model; specifically, assess offenders’ risk of recidivism, with more restrictive and intensive efforts focused on high-risk offenders; match treatment and rehabilitation efforts to offenders’ individual criminogenic needs (as identified in the risk assessment process) and deliver them in a way that is responsive to their individual learning style, motivation, personality and strengths. This will require regular review of staff performance. How performance, as well as fidelity, will be measured should be detailed in a comprehensive program evaluation plan established *prior to* implementation.

## BIBLIOGRAPHY

- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgement project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34*, 341-382.
- Anderson, D. A. (1999). The aggregate burden of crime. *Journal of Law and Economics, 42*, 611-642.
- Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency, 52*, 7-27.
- Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice, 15*, 346-360.
- Bonta, J., & Andrews, D. A. (2007). *Risk-need-responsivity model for offender assessment and rehabilitation* (User Report 2007-06). Ottawa, Ontario: Public Safety Canada.
- Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123*, 123-142.
- Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics – Simulation and Computation, 29*, 860-864.
- Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical And Experimental Neuropsychology, 23*, 695-700.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict in patient aggression: Reliability and validity of START assessments. *Psychological Assessment, 24*, 685-700.
- Desmarais, S. L., Van Dorn, R. A., Telford, R. P., Petrila, J., & Coffey, T. (2012). Characteristics of START assessments completed in mental health jail diversion programs. *Behavioral Sciences & the Law, 30*, 448-469.
- Douglas, K. S., Otto, R., Desmarais, S. L., & Borum, R. (in press). Clinical forensic psychology. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology, volume 2: Research methods in psychology*. Hoboken, NJ: John Wiley & Sons.

- Douglas, K. S., Skeem, J. L., & Nicholson, E. (2011). Research methods in violence risk assessment. In B. Rosenfeld & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 325-346). Hoboken, NJ: John Wiley & Sons, Inc.
- Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). The prediction of violence and antisocial behaviour: A systematic review and meta-analysis of the utility of risk assessment instruments in 73 samples involving 24,827 individuals. *British Medical Journal*, *345*, e4692.
- Federal Bureau of Investigation (FBI). (2012). *Crime in the United States, 2011*. Washington, D.C.: Authors.
- Gendreau, P., Goggin, C., & Little, T. (1996). *Predicting adult offender recidivism: What works!* (Cat. No. JS4-1/1996-7E). Ottawa, ON: Public Works and Government Services Canada.
- Glaze, L. E. (2011). *Correctional population in the United States, 2011*. Washington, D.C.: Bureau of Justice Statistics.
- Hanson, R. K., & Harris, A. J. R. (2000). A structured approach to evaluating change among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *13*, 105-122.
- Harris, G. T., Rice, M. E., & Quinsey, V. L. (2010). Allegiance or fidelity? A clarifying reply. *Clinical Psychology: Science and Practice*, *17*, 82-89.
- Hart, S. D., Michie, C., & Cooke, D. (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *The British Journal Of Psychiatry*, *190*(Suppl 49), s60-s65.
- Hart, S. D., Webster, C. D., & Douglas, K. S. (2001). Risk management using the HCR-20: A general overview of focusing on historical factors. In K. S. Douglas, C. D. Webster, S. D. Hart, D. Eaves, & J. R. P. Ogloff (Eds.), *HCR-20 violence risk management companion guide* (pp. 27-40). Burnaby, Canada/Tampa, FL: Simon Fraser University, Mental Health, Law & Policy Institute/University of South Florida, Dept. of Mental Health Law & Policy.
- Heilbrun, K., Dvoskin, J., Hart, S., & McNiel, D. (1999). Violence risk communication: Implications for research, policy, and practice. *Health, Risk & Society*, *1*, 91-105,
- Holtfreter, K., & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women. *Journal Of Contemporary Criminal Justice*, *23*, 363-382.
- Kyckelhahn, T. (2012). *Justice Expenditure And Employment Extracts, 2007 - Revised*. Washington, D.C.: Bureau of Justice Statistics.
- Langan, P. A. & Levin, D. J. (2002). *Recidivism of prisoners released in 1994* (NCJ 193427). Washington, D.C.: Bureau of Justice Statistics.

Leistico, A. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32*, 28-45.

Liptak, A. (2008, April 23). Inmate count in U.S. dwarfs other nations. *The New York Times*. Retrieved from <http://www.nytimes.com>.

Lowenkamp, C. T., Pealer, J., Smith, P., & Latessa, E. J. (2006). Adhering to the risk and need principles: Does it matter for supervision-based programs? *Federal Probation, 70*, 3-8.

Mamalian, C. A. (2011). *State of the science of pretrial risk assessment*. Washington, D.C.: Bureau of Justice Assistance.

Pew Center on the States (2009). *One in 31: The long reach of American corrections*. Washington, DC: The Pew Charitable Trusts.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*, 615-620.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences and the Law, 31*, 8-22.

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in studies of risk assessment instruments: A second-order systematic review. *Behavioral Sciences & the Law, 31*, 55-73.

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice & Behavior, 37*, 965-988.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science, 20*, 38-42.

Smith, P., Cullen, F., & Latessa, E. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy, 8*, 183-208.

Vincent, G. M., Guy, L. M., & Grisso, T. (2012). *Risk assessment in juvenile justice: A guidebook for implementation*. John D. And Catherine T. MacArthur Foundation. Available at: <http://modelsforchange.net/publications/346>

Walmsley, R. (2010). *World prison population list, 9th edition*. London: International Centre for Prison Studies.

Walters, G. D. (2003). Outcomes with the Psychopathy Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences & the Law, 21*, 89-102.

Walters, G. D. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice & Behavior*, 33, 279-304.

Wilson, C. M., Desmarais, S. L., Nicholls, T. L., Hart, S. D., & Brink, J. (in press). Incremental validity of dynamic factors in the assessment of violence risk. *Law and Human Behavior*.



## APPENDIX A

### *List of Jurisdiction-Specific Risk Assessment Instruments*

1. Alabama Risk and Needs Assessment
2. Allegheny County Risk Assessment
3. Arizona Risk Assessment Suite
4. Arkansas Post-Prison Board Transfer Risk Assessment
5. California Parole Violation Decision Making Instrument
6. California Static Risk Assessment
7. Colorado Actuarial Risk Assessment Scale
8. Connecticut Salient Factor Score
9. Delaware Parole Board Risk Assessment
10. Georgia Board of Pardons and Parole's Field Log of Interaction Data
11. Georgia Parole Behavior Response and Adjustment Guide
12. Georgia Parole Decisions Guidelines Grid System
13. Georgia Department of Corrections Offender Tracking Information System
14. Hawaii Risk and Needs Assessment
15. Illinois Risk Assessment Instrument
16. Illinois Risks, Assets and Needs Assessment Tool
17. Indiana Risk Assessment System
18. Kentucky Pretrial Risk Assessment Instrument
19. Kentucky Parole Guidelines Risk Assessment Instrument
20. Iowa Board of Parole Risk Assessment
21. Louisiana Risk Needs Assessment
22. Maryland Public Safety Risk Assessment
23. Michigan Parole Guidelines Score Sheet
24. Mississippi Parole Risk Instrument
25. Missouri Sentencing Assessment Risk Instrument
26. Missouri Parole Board Salient Factor Guidelines
27. Montana Risk Assessment Instrument
28. Nebraska Criminal History Assessment instrument
29. Nevada Parole Risk Assessment

30. New Mexico Risk and Needs Assessment
31. North Carolina Risk Needs Assessment
32. Oregon Criminal History/Risk Assessment
33. Public Safety Checklist for Oregon
34. Orange County Pretrial Risk Assessment
35. Rhode Island Parole Risk Assessment
36. South Carolina Parole Risk Assessment Instrument
37. South Dakota Initial Community Risk/Needs Assessment
38. State of Hawaii LSI-R Proxy
39. Tennessee Offender Risk Assessment/Needs Assessment
40. Tennessee Parole Grant Prediction Scale and Guidelines
41. Texas Parole Risk Assessment Instrument
42. Utah Criminal History Assessment
43. Vermont Parole Board Risk Assessment
44. Virginia Pretrial Risk Assessment Instrument
45. Virginia Risk Assessment Tool
46. Washington Risk Level Classification
47. West Virginia Parole Board Assessment

## APPENDIX B

### *Glossary of Terms*

#### *Actuarial Risk Assessment*

Mechanical approach to risk assessment in which offenders are scored on a series of items statistically associated with recidivism risk in the sample of offenders upon whom the instrument was developed. The total score is cross-referenced with a statistical table that translates the score into an estimate of recidivism risk during a specified timeframe.

#### *Area Under the Curve (AUC)*

Performance indicator measuring the probability that a randomly selected offender who recidivated during follow-up would have received a higher risk classification using a given risk assessment approach than a randomly selected offender who did not recidivate during follow-up.

#### *Cohen's d*

Performance indicator measuring the standardized mean difference between the estimated level of risk or total score of offenders who did and did not recidivate during follow-up.

#### *Dynamic Factor*

Changeable characteristics (e.g., substance abuse) that establish a relative level of risk and help inform intervention; they can be either relatively *stable*, changing relatively slowly over time (e.g., antisocial cognition) or *acute*, changing more quickly over time (e.g., mood state).

#### *Kappa (k)*

Measure of inter-rater reliability representing the percentage of categorizations (e.g., low, moderate or high risk) upon which multiple assessors agreed, statistically corrected for chance.

#### *Intra-Class Correlation Coefficient (ICC)*

Measure of inter-rater reliability representing the strength of agreement between multiple assessors on *continuous* variables (e.g., total scores), statistically corrected for chance.

#### *Meta-analysis*

*Systematic review* that includes a quantitative synthesis of the findings of *primary research*.

#### *Observed Agreement*

Measure of inter-rater reliability representing the percentage of categorizations (e.g., low, moderate or high risk) upon which multiple assessors agreed.

### *Odds ratio (OR)*

Performance indicator measuring the odds of the risk estimate in an offender who recidivates during follow-up being one higher than the risk estimate of an offender who does not recidivate.

### *Parole*

Conditional release of a prisoner before the expiration of his or her sentence subject to conditions supervised by a designated parole officer.

### *Performance Indicator*

Statistical measure of predictive validity.

### *Point-Biserial Correlation Coefficient ( $r_{pb}$ )*

Performance indicator measuring the direction and strength of the association between a *continuous predictor* (e.g., total score) and a *dichotomous outcome* (e.g., recidivating vs. not).

### *Primary Research*

Collection of new data that does not already exist.

### *Probation*

Release of an offender from detention or sentence served in the community in lieu of detention, subject to conditions supervised by a probation officer.

### *Protective Factor*

Characteristic of the offender (e.g., physical health, mental health, attitudes), his or her physical and/or social environment (e.g., neighborhood, family, peers) or situation (e.g., living situation) that is associated with a decrease in the likelihood of offending.

### *Recidivism*

Relapse into criminal behavior by an individual who has previously been convicted of one or more offenses.

### *Risk Assessment*

Process of estimating the likelihood an offender will recidivate to identify those at higher risk and in greater need of intervention. Also may assist in the identification of treatment targets and the development of risk management and treatment plans.

### *Risk Assessment Instrument*

Instrument composed of empirically- or theoretically-based risk and/or protective factors used to aid in the assessment of recidivism risk.

### *Risk Factor*

Characteristic of the offender (e.g., physical health, mental health, attitudes), his or her physical and/or social environment (e.g., neighborhood, family, peers) or situation (e.g., living situation) that is associated with an increase in the likelihood of offending.

### *Somer's d*

Performance indicator measuring the direction and strength of the association between an *ordinal predictor* (e.g., estimate of risk as low, moderate or high) and a *dichotomous outcome* (e.g., recidivating vs. not).

### *Structured Professional Judgment*

Structured approach to risk assessment focused on creating individualized and coherent risk formulations and comprehensive risk management plans. Assessors estimate risk through consideration of a set number of factors that are empirically and theoretically associated with the outcome of interest. Total scores are not used to make the final judgments of risk. Instead, assessors consider the relevance of each item to the individual offender, as well as whether there are any case specific factors not explicitly included in the list.

### *Static Factor*

A historical or otherwise unchangeable characteristics (e.g., history of antisocial behavior) that help establish absolute level of risk.

### *Systematic Review*

A process in which the empirical literature from multiple primary studies on a particular topic meeting pre-determined inclusion and exclusion criteria is descriptively analyzed.

### *Technical Violation*

A breach of the conditions of parole or probation.

### *Unstructured Risk Assessment*

A subjective assessment of recidivism risk based on the assessor's intuition, knowledge of theory, and professional experience.

# RISK-NEEDS ASSESSMENT: CONSTITUTIONAL AND ETHICAL CHALLENGES

Melissa Hamilton\*

## TABLE OF CONTENTS

I.	INTRODUCTION . . . . .	231
II.	RISK-NEEDS INSTRUMENTS . . . . .	233
	A. <i>Utility of Risk-Needs Data</i> . . . . .	234
	B. <i>Evolution of Risk-Needs Tools</i> . . . . .	236
III.	CRITICAL OBSERVATIONS OF RISK-NEEDS ASSESSMENTS . . . . .	240
	A. <i>Constitutional Considerations</i> . . . . .	242
	1. <i>Equal Protection</i> . . . . .	242
	a. <i>Rational Basis Review</i> . . . . .	243
	b. <i>Heightened Review: Gender</i> . . . . .	250
	c. <i>Strict Scrutiny: Race, Alienage, and Fundamental Rights</i> . . . . .	256
	d. <i>Proxies</i> . . . . .	261
	2. <i>Prisoners' Rights</i> . . . . .	263
	a. <i>Fundamental Rights</i> . . . . .	264
	b. <i>Due Process</i> . . . . .	267
	3. <i>Sentencing</i> . . . . .	271
	B. <i>Ethical and Normative Concerns</i> . . . . .	277
IV.	THE FUTURE OF SOCIODEMOGRAPHIC FACTORS . . . . .	280
V.	CONCLUSIONS . . . . .	285
	APPENDIX A: POPULAR RISK ASSESSMENT TOOLS . . . . .	286

## I. INTRODUCTION

Evidence-based practices are in vogue as the post-modern savior within criminal justice. Not long ago, a legal commentator observed that risk analysis dominated the law in the areas of environmental, health, and safety issues but had not yet become established in criminal law and procedure.<sup>1</sup> Whatever the validity of the statement at the time, across jurisdictions the criminal justice system has

---

\* Visiting Criminal Law Scholar, University of Houston Law Center; J.D., The University of Texas School of Law; Ph.D, The University of Texas at Austin. © 2014, Melissa Hamilton.

1. Jonathan Remy Nash, *The Supreme Court and the Regulation of Risk in Criminal Law Enforcement*, 92 B.U. L. REV. 171, 173 (2012).

embraced the evidence-based practices movement.<sup>2</sup> The United States' economic ills and its record-breaking rate of incarceration have convinced policymakers to adopt new strategies to constrain a dependence on imprisonment, encourage alternative rehabilitative programming, reduce recidivism risk, and improve public safety.<sup>3</sup> The evidence-oriented model utilizes the best data available from the empirical sciences to identify and classify individuals based on their potential future risk of reoffending, and then to manage offender populations accordingly.

Well-informed decisions are critical to achieving a proper balance among such interests as protecting the public and efficiently expending government resources, while at the same time respecting individuals' liberty interests.<sup>4</sup> The ideology of risk is now considered at the heart of such a balancing act in that information about a defendant's risk of recidivism informs an expanding number and variety of criminal justice decisions.<sup>5</sup> Interested observers have referred to risk-based philosophies as promoting a "preventive, future-oriented logic of risk,"<sup>6</sup> representing "risk factorology,"<sup>7</sup> and embracing a stance toward risk aversion.<sup>8</sup>

The assessment of risk cannot constitute a simplistic enterprise as human behavior is often capricious. Advocates of the new risk penology properly continue to search for improvements in risk assessment practices by incorporating scientific advances from interdisciplinary research fields.<sup>9</sup> Empirical studies influenced a more recent revolution of the risk penology toward the risk-needs model, which adds to the prediction of future risk a framework for engaging principles of effective correctional interventions addressing criminogenic needs.<sup>10</sup> To be sure, academics across disciplines have long been studying criminal offending. The idea that criminal justice should not be simply focused on warehousing offenders was

---

2. Christopher T. Lowenkamp et al., *When a Person Isn't a Data Point: Making Evidence-Based Practice Work*, 76 FED. PROBATION, no. 3, 2012, at 11, 12.

3. Douglas A. Berman, *Re-Balancing Fitness, Fairness, and Finality for Sentences*, 4 WAKE FOREST J. L. & POL'Y 151, 172–73 (2014); David Dagan & Steven M. Teles, *Locked In? Conservative Reform and the Future of Mass Incarceration*, 651 ANNALS AM. ACAD. POL. & SOC. SCI. 266, 270–71 (2014).

4. Michael L. Rich, *Limits on the Perfect Preventive State*, 46 CONN. L. REV. 883, 932–33 (2014); Jay P. Singh, *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. & L. 55, 55 (2013).

5. Lowenkamp et al., *supra* note 2, at 12–13; Christopher Slobogin, *A Jurisprudence of Dangerousness*, 98 NW. U. L. REV. 1, 1 (2003) ("Dangerousness determinations permeate the government's implementation of its police power.").

6. Mariana Valverde et al., *Legal Knowledges of Risk*, in LAW AND RISK 86, 116 (Law Comm'n of Canada ed., 2005).

7. Hazel Kemshall, *Crime and Risk*, in RISK IN SOCIAL SCIENCE 76, 82 (Peter Taylor-Goodby & Jens O. Zinn eds., 2006).

8. Min Yang et al., *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. BULL. 740, 740 (2010).

9. Jennifer L. Skeem & John Monahan, *Current Directions in Violence Risk Assessment*, 20 CURRENT DIRECTIONS IN PSYCHOL. SCI. 38, 41 (2011).

10. See Jan Looman & Jeffrey Abracen, *The Risk Need Responsivity Model of Offender Rehabilitation: Is There Really a Need for a Paradigm Shift?*, 8 INT'L J. BEHAV. CONSULTATION & THERAPY 30, 30 (2013).

represented with a zeitgeist-like emphasis on rehabilitation until the 1970s.<sup>11</sup> However, since that time the American criminal justice system dramatically deviated away from rehabilitative goals toward retribution, which helps account for the high rate of incarceration ever since. Nonetheless, a current of disappointment with the high financial and social costs of over-incarceration have convinced officials across the country to explore new ideologies by considering alternatives to incarceration and the adoption of practices that work to reduce recidivism rates by addressing criminogenic needs. In sum, this so-called “neorehabilitation” model—meaning the rehabilitation of rehabilitation—seeks to improve upon past practices by incorporating evidence-based practices.<sup>12</sup> Despite good intentions, controversies emerged.

This Article proceeds as follows. Section II surveys the variety of criminal justice decisions currently informed by risk-needs assessment and introduces several of the most popular tools. Section III reviews constitutional and moral objections to risk-needs tools, such as those recently raised by Attorney General Eric Holder, targeting a host of sensitive factors contained therein, such as demographic and other immutable characteristics. The constitutional analysis engages equal protection, prisoners’ rights, due process, and sentencing law. The text also examines the philosophical polemic aimed uniquely at sentencing as to whether risk should play any role in determining punishment. Neorehabilitation is not necessarily always the golden standard. Across criminal justice decisions, punishment theories variously involve sometimes conflicting perspectives depending on whether officials are reliant upon retribution, deterrence, incapacitation, and/or rehabilitation as the orienting value system(s). The utility of risk-needs considerations likewise will vary by the prevailing punishment philosophy. Section IV appraises potential alternatives for risk-needs methodologies if the concerns so raised prove legitimate. Any option comes with significant consequences. Retaining offensive variables incites political and ethical reproach, while simply removing them weakens statistical validity of the underlying models and diminishes the promise of evidence-based practices. With respect to sentencing, promoting an emphasis on risk diminishes the focus of punishment on blameworthiness, while neglecting risk and needs serves to sabotage a core objective of the contemporary neorehabilitation model of harnessing the ability to identify and divert low risk offenders to community-based alternatives offering culturally-sensitive rehabilitative services. Section V concludes.

## II. RISK-NEEDS INSTRUMENTS

The employment of automated tools that capitalize on the ideology of risk is enjoying its heyday in criminal justice. Numerous scholars and scientists have

---

11. Berman, *supra* note 3, at 158.

12. Jessica M. Eaglin, *Against Neorehabilitation*, 66 SMU L. REV. 189, 193 (2013).



hastened to develop and cross-validate a variety of tools. Risk-needs assessment has become a competitive industry with governmental and for-profit businesses issuing a host of instruments that are either generic in nature or targeted to specific groups (e.g., youth, mentally disordered) or offense types (e.g., sex offenders, violent aggressors).<sup>13</sup> “Recidivism prediction is ubiquitous. Everybody’s doing it. There is an enormous academic and professional literature. Unprecedented private sector involvement has occurred in designing and marketing instruments and providing services to government.”<sup>14</sup> Some of the tools are proprietary and require payment for their use, while others are in the public domain.<sup>15</sup> Officials in the criminal justice system have become convinced that predicting risk and addressing criminogenic needs are crucial to the core goals in criminal justice of protecting the public, securing correctional institutions, reducing recidivism, providing rehabilitative programming, and at the same time saving resources.

#### A. *Utility of Risk-Needs Data*

A justification for the prevalence of risk-based datasets and models is the growth in the number and type of decisions for which they are perceived to be useful. Initially, evidence-based practices were adopted to inform post-conviction decisions and management strategies, such as parole determinations,<sup>16</sup> supervised release conditions, provision of reentry services,<sup>17</sup> decisions to revoke supervision, and judgments concerning probation and parole sanctions.<sup>18</sup> Risk analysis is helpful in crafting release conditions as studies indicate overly burdensome restrictions can harm many otherwise low risk offenders.<sup>19</sup> The adoption of the evidence-based model in general, and the implementation of risk-needs tools more specifically, has recently been promoted in pretrial contexts,<sup>20</sup> such as pretrial

---

13. Leon Neyfakh, *You Will Commit a Crime in the Future: Inside the New Science of Predicting Violence*, BOSTON GLOBE, Feb. 20, 2011, [http://www.boston.com/bostonglobe/ideas/articles/2011/02/20/you\\_will\\_commit\\_a\\_crime\\_in\\_the\\_future/](http://www.boston.com/bostonglobe/ideas/articles/2011/02/20/you_will_commit_a_crime_in_the_future/).

14. Michael Tonry, *Legal and Ethical Issues in the Prediction of Recidivism*, 26 FED. SENT’G REP. 167, 167 (2014).

15. Susan Turner & Julie Gerlinger, *Risk Assessment and Realignment*, 53 SANTA CLARA L. REV. 1039, 1045 (2013).

16. David DeMatteo et al., *Investigating the Role of the Psychopathy Checklist-Revised in United States Case Law*, 20 PSYCHOL. PUB. POL’Y & L. 96, 96–97, 100 (2014).

17. *Barge v. Pa. Bd. of Prob. & Parole*, 39 A.3d 530, 549 (Pa. Commw. Ct. 2012).

18. PEW CTR. ON THE STATES, RISK/NEEDS ASSESSMENT 101: SCIENCE REVEALS NEW TOOLS TO HELP MANAGE OFFENDERS 2 (2011), available at [http://www.pewtrusts.org//media/legacy/uploadedfiles/pcs\\_assets/2011/PewRiskAssessmentbriefpdf.pdf](http://www.pewtrusts.org//media/legacy/uploadedfiles/pcs_assets/2011/PewRiskAssessmentbriefpdf.pdf).

19. Timothy P. Cadigan & Christopher T. Lowenkamp, *Preentry: The Key to Long-Term Criminal Justice Success?*, 75 FED. PROBATION no. 2, 2011, at 74, 74.

20. *Id.* at 74–75.

diversion,<sup>21</sup> deferred adjudication, bail, and plea negotiations,<sup>22</sup> and juvenile transfers to adult court.<sup>23</sup> The concern now is not just an immediate interest in reserving pretrial detention for those likely to fail if released into the community, but also a longer term recognition that pretrial incarceration positively correlates with post-conviction failures.<sup>24</sup> Risk-needs assessments are immensely popular for a variety of similar decisions made in specialty courts, such as drug courts<sup>25</sup> and reentry courts.<sup>26</sup>

Empirically-based evaluations of future recidivism risk and criminogenic needs are also helpful in other management circumstances, such as designation as a sexually violent predator for purposes of civil commitment,<sup>27</sup> sex offender registration classification,<sup>28</sup> inmate security classification levels, institutional placement,<sup>29</sup> and therapy options in treatment.<sup>30</sup> Perhaps the most recent legal arena to turn to risk-needs is sentencing. The idea being to guide sentencers in distinguishing high-risk defendants, for whom preventive incapacitation—perhaps even the death penalty—may be suitable, from low-risk offenders who may fittingly be diverted from prison.<sup>31</sup> Risk-needs data also are informing sentencing decisions in the consideration of suitable alternatives to prison and tailoring conditions of community confinement to individual and cultural circumstances.<sup>32</sup>

Experts maintain there exists a “central eight” risk-needs categories that research consistently show are most associated with recidivism.<sup>33</sup> Comprising the central eight, the “big four” are antisocial attitudes, antisocial associates, antisocial personalities, and criminal history, while the “moderate four” include substance abuse, family characteristics, education and employment, and lack of prosocial

---

21. Joseph M. Zlatic et al., *Pretrial Diversion: The Overlooked Pretrial Services Evidence-Based Practice*, 74 FED. PROBATION, no. 1, 2010, at 28, 33.

22. MAREA BEEMAN & AIMEE WICKMAN, THE JUSTICE MGMT. INST., RISK AND NEEDS ASSESSMENT 3 (2013), available at <https://www.yumpu.com/en/document/view/23676461/risk-needs-assessment-justice-management-institute/1>.

23. Michael J. Vitacco et al., *The Role of the Violence Risk Appraisal Guide and Historical, Clinical, Risk-20 in U.S. Courts*, 18 PSYCHOL. PUB. POL'Y & L. 361, 381 (2012).

24. Cadigan & Lowenkamp, *supra* note 19, at 74–75.

25. Max Deitchler, *You Can't Manage what you Don't Measure: An Evaluation of Arkansas's Drug Courts*, 64 ARK. L. REV. 715, 735 (2011).

26. DEBBIE BOAR & CHRISTOPHER WATLER, CENTER FOR COURT INNOVATION, REENTRY COURT TOOLKIT: A GUIDE FOR REENTRY COURT PRACTITIONERS 5 (2012), available at [http://www.courtinnovation.org/sites/default/files/documents/reentry\\_toolkit.pdf](http://www.courtinnovation.org/sites/default/files/documents/reentry_toolkit.pdf).

27. DeMatteo et al., *supra* note 16, at 96–97.

28. Doe v. Sex Offender Registry Bd., 4 N.E.3d 1264, 1269 (Mass. App. Ct. 2014).

29. PEW CTR. ON THE STATES, *supra* note 18, at 2.

30. Brooks v. Roy, 881 F. Supp. 2d 1034, 1043 (D. Minn. 2012).

31. DeMatteo et al., *supra* note 16, at 100; Nancy J. King, *Sentencing and Prior Convictions: The Past, the Future, and the End of the Prior-Conviction Exception to Apprendi*, 97 MARQ. L. REV. 523, 540–41 (2014) (citing statutes and the Model Penal Code revised).

32. PEW CTR. ON THE STATES, *supra* note 18, at 1–2.

33. Julienne James et al., *A View from the States: Evidence-Based Public Safety Legislation*, 102 J. CRIM. L. & CRIMINOLOGY 821, 825 (2012).

leisure or recreation (though it is recognized that the moderate four largely influence recidivism via the big four).<sup>34</sup> Thus, risk-needs instruments in the field of criminal offending often embed at least a few factors from the central eight categories.<sup>35</sup>

The utility of using risk-needs instruments has attracted energetic support from many reputable policy centers, namely the Justice Center of the Council of State Governments,<sup>36</sup> the Justice Management Institute,<sup>37</sup> the Center for Effective Public Policy,<sup>38</sup> the Vera Institute,<sup>39</sup> and the Center for Court Innovation.<sup>40</sup> Loyalty to evidence-based corrections is equally evident at the state and local levels. For instance, the Judicial Branch of California officially labels the implementation of evidence-based practices in sentencing and corrections policy and practice as “perhaps the most important reform” in criminal justice.<sup>41</sup> The New York City Department of Probation likewise proclaims that it “is in the midst of incorporating evidence-based policies and practices into virtually everything [they] do.”<sup>42</sup>

### B. Evolution of Risk-Needs Tools

The instruments at the heart of evidence-based corrections practices have evolved over time such that a historical perspective unveils four generations of assessment tools. The first generation of assessments consisted of clinicians conducting unstructured or semi-structured interviews to extract relevant information that, based on the professional’s experience and knowledge, constituted recidivism risk factors.<sup>43</sup> First generation assessment methodologies formed the basis for modern risk assessment practices; although they have largely been

---

34. Michael S. Caudy et al., *How Well Do Dynamic Needs Predict Recidivism? Implications for Risk Assessment and Risk Reduction*, 41 J. CRIM. JUST. 458, 459 (2013).

35. See *infra* app. A.

36. COUNCIL OF STATE GOV’TS, LESSONS FROM THE STATES: REDUCING RECIDIVISM AND CURBING CORRECTIONS COSTS THROUGH JUSTICE REINVESTMENT 6–7 (2013) [hereinafter LESSONS FROM THE STATES], available at <http://csgjusticecenter.org/jr/publications/lessons-from-the-states>.

37. BEEMAN & WICKMAN, *supra* note 22, at 3.

38. *Evidence-Based Decision Making*, CTR. FOR EFFECTIVE PUB. POL’Y, <http://cepp.com/evidence-based-practice> (last visited Nov. 16, 2014).

39. VERA INST. OF JUSTICE AND DELAWARE JUSTICE REINVESTMENT TASK FORCE 1–2 (Oct. 12, 2011), available at [http://ltgov.delaware.gov/taskforces/djrtf/DJRTF\\_Risk\\_Assessment\\_Memo.pdf](http://ltgov.delaware.gov/taskforces/djrtf/DJRTF_Risk_Assessment_Memo.pdf) [hereinafter VERA MEMORANDUM].

40. MICHAEL REMPEL, CTR. FOR COURT INNOVATION, EVIDENCE-BASED STRATEGIES FOR WORKING WITH OFFENDERS 1–2 (2014), available at <http://www.courtinnovation.org/research/evidence-based-strategies-working-offenders>.

41. *Evidence-Based Practice*, THE JUDICIAL BRANCH OF CALIF., <http://www.courts.ca.gov/5285.htm> (last visited Nov. 16, 2014).

42. *Evidence-Based Policies and Practices*, NEW YORK CITY DEP’T OF PROB., <http://www.nyc.gov/html/prob/html/about/evidence.shtml> (last visited Nov. 16, 2014).

43. Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 CRIM. JUST. & BEHAV. 1095, 1095 (2008).

supplanted by later generation tools because of perceived improvements in predictive validity.<sup>44</sup>

Second generation assessments were empirically-based scoring instruments of those variables that were statistically shown to correlate with recidivism.<sup>45</sup> The focus of second generation instruments was on risk rather than rehabilitation needs, and they were intended to be brief and efficiently scored.<sup>46</sup> Examples of second generation instruments are the Violence Risk Appraisal Guide (VRAG),<sup>47</sup> Static-99,<sup>48</sup> and the federal Pre-Trial Risk Assessment tool (PTRA). VRAG remains the most popular tool to assess violent recidivism and contains twelve factors, such as age, marital status, criminal history, and psychopathy.<sup>49</sup> Static-99 is most widely used for sexual recidivism and contains ten static factors, five of which relate to criminal history, with several variables respecting victim type, plus age and cohabitation history.<sup>50</sup> A more recently created instrument, though it still falls within the second generation genre, is the federal probation office's PTRA tool. PTRA rates eleven items, including the seriousness of the current charge, education, home ownership, and citizenship.<sup>51</sup>

The third generation's scientific advancements combined actuarial assessment with directed professional judgment and integrated static with dynamic factors.<sup>52</sup> Static risk factors normally are historical, unchangeable, and generally not amenable to interventions.<sup>53</sup> Dynamic factors incorporate criminogenic needs, which are often mutable in nature and, therefore, may be proper targets for rehabilitative programming.<sup>54</sup> The HCR-20 is a structured professional judgment guide for violence risk assessment and management.<sup>55</sup> Its developer recently

---

44. See Tim Brennan et al., *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21, 21–22 (2009) (noting the first generation “approach relied on clinical and professional judgment in the absence of any explicit or objective scoring rules”).

45. Fass et al., *supra* note 43, at 1095–96.

46. Brennan et al., *supra* note 44, at 22.

47. Debra A. Pinals et al., *Violence Risk Assessment*, in *SEX OFFENDERS: IDENTIFICATION, RISK ASSESSMENT, TREATMENT, AND LEGAL ISSUES* 49, 55 (Fabian M. Saleh et al. eds., 2009).

48. Georgia D. Barnett & Ruth E. Mann, *Good Lives and Risk Assessment: Collaborative Approaches to Risk Assessment with Sexual Offenders*, in *GOOD PRACTICE IN ASSESSING RISK: CURRENT KNOWLEDGE, ISSUES AND APPROACHES* 139, 140 (Hazel Kemshall & Bernadette Wilkinson eds., 2011).

49. Skeem & Monahan, *supra* note 9, at 39.

50. R. Karl Hanson & David Thornton, *Improving Risk Assessments for Sex Offenders: A Comparison of Three Actuarial Scales*, 24 LAW & HUM. BEHAV. 119, app. I (2000).

51. Timothy P. Cadigan et al., *The Re-validation of the Federal Pretrial Services Risk Assessment (PTRA)*, 76 FED. PROBATION, no. 2, 2012, at 3, 6.

52. Fass et al., *supra* note 43, at 1095–96.

53. *Id.* at 1096.

54. Paul Gendreau et al., *A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!*, 34 CRIMINOLOGY 575, 575–76 (1996).

55. KEVIN S. DOUGLAS ET AL., *HCR-20 VIOLENCE RISK ASSESSMENT SCHEME: OVERVIEW AND ANNOTATED BIBLIOGRAPHY* 6 (2014), available at <http://kdouglas.files.wordpress.com/2014/01/hcr-20-annotated-bibliography-version-12-january-20142.pdf>.

claimed that the HCR-20 is among the world's most widely used and best validated risk-needs instruments for violent reoffending.<sup>56</sup> In summary, the

HCR-20 is so-named for its inclusion of 20 risk factors in Historical, Clinical, and Risk management domains. The instrument contains 10 historical, largely static, risk factors that fall into three general categories (problems in adjustment or living, problems with mental health, and past antisocial behavior) and 10 potentially changeable, dynamic risk factors. Five of these concern current clinical status such as negative attitudes and active symptoms of major mental illness (the Clinical scale), and five concern future situational risk factors such as lack of plan feasibility and treatment noncompliance (the Risk Management scale).<sup>57</sup>

The Level of Service Inventory-Revised (LSI-R), also a third generation tool,<sup>58</sup> is likewise a structured professional judgment instrument and is, according to a national survey by the Vera Institute, the most commonly used generic risk-needs tool across American criminal justice agencies.<sup>59</sup>

[The LSI-R] contains 54 items rationally grouped according to the following 10 subcomponents representing different risk/need areas: Criminal History, Education/Employment, Finances, Family/Marital, Accommodations, Leisure/Recreation, Companions, Alcohol/Drug, Emotional/Personal, and Attitude/Orientation. Items are scored as either present or absent, based on a semistructured interview and review of available file information, and subsequently summed to yield a total score. Higher scores reflect a greater risk of recidivism and need for intervention.<sup>60</sup>

In the latest iteration, fourth generation assessments supplemented the risk-needs combination with responsivity principles and a longer perspective on case management spanning from intake through case closure.<sup>61</sup> “Responsivity is defined as tailoring case plans to the individual characteristics, circumstances, and learning style of each offender.”<sup>62</sup> Fourth generation tools are often automated with technological applications using algorithmic scoring. The federal probation system developed its Post Conviction Risk Assessment (PCRA) as a fourth

---

56. *Id.* at 3.

57. Laura S. Guy et al., *Assessing Risk of Violence Using Structured Professional Judgment Guidelines*, 12 J. FORENSIC PSYCHOL. PRAC. 270, 272 (2012).

58. Pinals et al., *supra* note 47, at 56.

59. VERA MEMORANDUM, *supra* note 39, at 4.

60. David J. Simourd & P. Bruce Malcolm, *Reliability and Validity of the Level of Service Inventory-Revised Among Federally Incarcerated Sex Offenders*, 13 J. INTERPERSONAL VIOLENCE 261, 264 (1998).

61. Fass et al., *supra* note 43, at 1096.

62. WINNIE ORE & CHRIS BAIRD, NAT'L COUNCIL ON CRIME & DELINQUENCY, BEYOND RISK AND NEEDS ASSESSMENTS 8 (2014), available at [http://nccdglobal.org/sites/default/files/publication\\_pdf/beyond-risk-needs-assessments.pdf](http://nccdglobal.org/sites/default/files/publication_pdf/beyond-risk-needs-assessments.pdf).

generation, software-based tool.<sup>63</sup> The PCRA scores a variety of static and dynamic factors, including education, employment, substance abuse, family problems, and procriminal attitudes.<sup>64</sup> The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is one of the best known fourth generation tools,<sup>65</sup> and is described as a “web-based tool designed to assess offenders’ criminogenic needs and risk of recidivism. Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision, and case management of offenders.”<sup>66</sup> Reflecting the progresses made in the fourth generation, COMPAS distinguishes itself:

Unlike other risk assessment instruments, which provide a single risk score, the COMPAS provides separate risk estimates for violence, recidivism, failure to appear, and community failure. In addition to the Overall Risk Potential, as represented by those four scales, the COMPAS provides a Criminogenic and Needs Profile for the offender. This profile provides information about the offender with respect to criminal history, needs assessment, criminal attitudes, social environment, and additional factors such as socialization failure, criminal opportunity, criminal personality, and social support.<sup>67</sup>

Overall, a confident synthesis of the proposed value of the current state of risk-needs tools is as follows:

Risk assessment tools now under consideration are more transparent, rely on data, and attempt to regularize th[e] instinct [to predict risk] and subject it to more scientifically rigorous examinations. Ensuring uniform application and the unbiased use of available data, these modern predictive tools are facilitated by the use of ‘structured, empirically-driven and theoretically driven’ instruments.<sup>68</sup>

The foregoing constitutes a rather brief introduction to the evolution of risk-needs tools and an identification of a few of the most popular in use today. The next section will provide a more extensive investigation of the application of risk-needs tools in criminal justice decisions, with a focus on constitutional law issues and moral considerations.

---

63. Christopher T. Lowenkamp et al., *The Federal Post Conviction Risk Assessment (PCRA): A Construction and Validation Study*, 10 PSYCHOL. SERVICES 87, 88 (2013).

64. James L. Johnson et al., *The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA)*, 75 FED. PROBATION, no. 2, 2011, at 16, 26 app. 2.

65. Fass et al., *supra* note 43, at 1097.

66. NORTHPOINTE, PRACTITIONER’S GUIDE TO COMPAS 1 (2013), available at [http://www.northpointeinc.com/files/technical\\_documents/FieldGuide2\\_012813.pdf](http://www.northpointeinc.com/files/technical_documents/FieldGuide2_012813.pdf).

67. Fass et al., *supra* note 43, at 1098.

68. Jordan M. Hyatt et al., *Reform in Motion: The Promise and Perils of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing*, 49 DUQ. L. REV. 707, 725 (2011) (citation omitted).

### III. CRITICAL OBSERVATIONS OF RISK-NEEDS ASSESSMENTS

The philosophy underlying evidence-based practices, along with its goal of informing a host of correctional decisions, certainly are laudable. Policymakers and justice officials should be praised for seeking out progressive ideas and engaging alternative options, as opposed to the frequent presumption of incarceration that has burdened their corrections systems over the last thirty years.<sup>69</sup> However, scholars and practitioners are debating the appropriateness of using risk-needs tools for criminal justice-oriented decisions due to the presence of potentially objectionable variables within them. Risk-needs tools incorporate a host of factors that are demographic in nature, score on measures involving personal and social functioning, increase risk predictions based on the presence of mental conditions and drug addictions, and rate attitudes indicative of an antisocial outlook. Consequently, a variety of the items scored in risk-needs assessments raise constitutional, ethical, and normative issues.<sup>70</sup> For reference, Appendix A contains a summary list of the factors and measures used in some of the most popular risk-needs instruments, sorted by generation.

Risk-needs tools normally score at least several demographic characteristics of the individuals evaluated. Among various instruments, these entail age,<sup>71</sup> gender,<sup>72</sup> citizenship,<sup>73</sup> and marital status.<sup>74</sup> Risk-needs tools orient toward rating demographic variables regarding various aspects of family of origin, including having lived with both biological parents until age sixteen,<sup>75</sup> a criminal family,<sup>76</sup> parental alcohol problem,<sup>77</sup> and current family situation.<sup>78</sup> Ratings are commonly provided relative to the individual's personal history, namely criminal background,<sup>79</sup> educational attainment,<sup>80</sup> and employment stability.<sup>81</sup> The instruments often contain

---

69. Melissa Hamilton, *Prison-by-Default: Challenging the Federal Sentencing Policy's Presumption of Incarceration*, 51 HOUS. L. REV. 1271, 1272–74 (2014) (noting that the United States earns world's highest incarceration rate).

70. Tonry, *supra* note 14, at 167, 169.

71. PCRA; PTRR; VRAG; Static-99; COMPAS. *See infra* app. A (summarizing risk assessment tools); NORTHPOINTE, *supra* note 66, at 20, 27 (2013); *see also* MINN. DEPT. OF CORRECTIONS, THE MINNESOTA SEX OFFENDER SCREENING TOOL-3.1 (MNSOST-3.1): AN UPDATE TO THE MNSOST-3, at 33 (2012) (describing MnSOST-3.1, a Minnesota sex offender screening tool).

72. COMPAS. *See infra* app. A (summarizing risk assessment tools); *see also* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 823 n.76 (2014) (listing instruments that incorporate gender). PCRA includes the Psychological Inventory of Criminal Thinking Styles (PICTS) with a gender-based scoring system.

73. PTRR. *See infra* app. A (summarizing risk assessment tools).

74. PCRA; VRAG. *See infra* app. A (summarizing risk assessment tools).

75. VRAG. *See infra* app. A (summarizing risk assessment tools).

76. LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools).

77. VRAG. *See infra* app. A (summarizing risk assessment tools).

78. PCRA; LSI-R. *See infra* app. A (summarizing risk assessment tools).

79. PCRA; PTRR; VRAG; Static-99; HCR-20; LSI-R. *See infra* app. A (summarizing risk assessment tools).

80. PCRA; PTRR; LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools).

81. PCRA; PTRR; HCR-20; LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools).

measures implicating socioeconomic status, such as financial condition,<sup>82</sup> ownership of home,<sup>83</sup> residential stability,<sup>84</sup> and living in a neighborhood with high crime<sup>85</sup> or illegal drug activity.<sup>86</sup>

Some risk-needs tools compile and rate various aspects of personal and social functioning. Examples consist of elementary school maladjustment<sup>87</sup> and problems with personal support,<sup>88</sup> in addition to factors focused on reliance on social services or public assistance,<sup>89</sup> which may suggest deficits in personal responsibility. Various measures rate relationship issues involving family, consisting of relationship with parents<sup>90</sup> and marital/family problems,<sup>91</sup> and social functioning, such as a history of problems with relationships,<sup>92</sup> social adjustment problems,<sup>93</sup> lack of pro social support,<sup>94</sup> and maintaining criminal acquaintances.<sup>95</sup>

Addictions and mental conditions are commonly integrated therein. These include problems with alcohol<sup>96</sup> or drugs,<sup>97</sup> a history of a mental disorder,<sup>98</sup> personality disorder,<sup>99</sup> psychopathy,<sup>100</sup> or of mental health treatment.<sup>101</sup> Several of the instruments judge attitudes, such as temperament towards supervision and change,<sup>102</sup> lack of insight,<sup>103</sup> personal instability,<sup>104</sup> and problems with stress and coping.<sup>105</sup>

Upon reviewing the foregoing summary, and the list of variables contained in Appendix A, one might well be both comforted that many of the factors appear

---

82. LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools).

83. PTR. *See infra* app. A (summarizing risk assessment tools).

84. LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools); *see also* EDWARD LATESSA, ET. AL, CREATION AND VALIDATION OF THE OHIO RISK ASSESSMENT SYSTEM, FINAL REPORT, 49 app. A (2009), *available at* [http://www.ocjs.ohio.gov/ORAS\\_FinalReport.pdf](http://www.ocjs.ohio.gov/ORAS_FinalReport.pdf) (describing the Ohio Risk Assessment System: Pretrial Assessment Tool).

85. LSI-R. *See infra* app. A (summarizing risk assessment tools).

86. LATESSA, ET. AL, *supra* note 84, at 49.

87. VRAG. *See infra* app. A (summarizing risk assessment tools); *see also* LSI-R (rating school suspensions and level of participation in school activities). *Id.*

88. HCR-20. *See infra* app. A (summarizing risk assessment tools).

89. LSI-R. *See infra* app. A (summarizing risk assessment tools).

90. PCRA; LSI-R. *See infra* app. A (summarizing risk assessment tools).

91. *Id.*

92. LSI-R; HCR-20. *See infra* app. A (summarizing risk assessment tools).

93. LSI-R; HCR-20; COMPAS. *See infra* app. A (summarizing risk assessment tools).

94. PCRA; LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools).

95. LSI-R; COMPAS. *See infra* app. A (summarizing risk assessment tools).

96. PCRA; VRAG; LSI-R. *See infra* app. A (summarizing risk assessment tools).

97. PTR; LSI-R; HCR-20; COMPAS. *See infra* app. A (summarizing risk assessment tools).

98. HCR-20; LSI-R. *See infra* app. A (summarizing risk assessment tools).

99. VRAG; HCR-20. *See infra* app. A (summarizing risk assessment tools).

100. *Id.*

101. LSI-R; HCR-20. *See infra* app. A (summarizing risk assessment tools).

102. PCRA; LSI-R; HCR-20. *See infra* app. A (summarizing risk assessment tools).

103. HCR-20. *See infra* app. A (summarizing risk assessment tools).

104. LSI-R; HCR-20. *See infra* app. A (summarizing risk assessment tools).

105. *Id.*



perfectly suited to assessing risk and criminogenic needs, yet likewise concerned that more than a few implicate—directly or by proxy—characteristics for which we are sensitive in terms of exploiting certain attributes to rate and classify individuals, perhaps even to punish. Therefore, reliance upon risk-needs assessments when they incorporate potentially problematic factors in the important arena of criminal justice decisions incites constitutional and moralistic concerns. The constitutional doctrines on point include equal protection, prisoners' rights, and sentencing law. The moral issues involve political unease when decisions are based on immutable characteristics over which individuals have no personal control or that may serve directly or by proxy to replicate discriminatory practices.

### *A. Constitutional Considerations*

By its nature, the use of risk-needs assessments to inform a host of correctional decisions animates several areas of relevant law. The most applicable constitutional guarantees encompass equal protection, prisoners' rights, due process, and rights in sentencing. This subsection will address each body of law as applied to risk-needs analysis in criminal justice decisionmaking, albeit recognizing these legal frameworks often overlap to some degree.

#### *1. Equal Protection*

The Equal Protection Clause embodies the philosophy that persons who are similarly situated ought to be treated alike.<sup>106</sup> The right exemplifies the central concept that individuals should be accorded fair treatment in the exercise of fundamental rights and that distinctions between groups based on impermissible criteria should be prohibited. Risk-needs instruments utilize a plethora of factors and characteristics to justify criminal justice decisions that may infringe upon fundamental rights or that differentiate between various groups with respect to benefits or burdens. Both results implicate equal protection issues. Regarding classifications, it should be noted that it is not always evident that any contrast in the treatment between groups normatively should be deemed unconstitutional. The Supreme Court has cautioned that equal protection's promise "must coexist with the practical necessity that most legislation classifies for one purpose or another, with resulting disadvantage to various groups or persons."<sup>107</sup>

The Supreme Court's development of the law of equal protection has resulted in three tiers of analysis: rational basis review, heightened review, and strict scrutiny. The vast majority of claims will fall within the lowest tier, typically the easiest test

---

106. *City of Cleburne v. Cleburne Living Ctr.*, 473 U.S. 432, 439 (1985). While the Fourteenth Amendment technically only applies to the states, the Supreme Court has ruled that its approach to equal protection claims pertains equally to the federal government via the Fifth Amendment's Due Process Clause. *Adarand Constructors v. Peña*, 515 U.S. 200, 217 (1995).

107. *Romer v. Evans*, 517 U.S. 620, 631 (1996).

for the government to win in sustaining its disparate treatment of a group. This first tier employs rational basis review, whereby the law or policy challenged will survive so long as it serves a legitimate public purpose and the classifications drawn are “reasonable in light of its purpose.”<sup>108</sup> The second tier requires a law or policy to receive heightened review if it either constructs classifications involving protected groups or infringes upon fundamental rights.<sup>109</sup> Heightened review involves either intermediate scrutiny or strict scrutiny. To date, the Supreme Court has only sanctioned gender<sup>110</sup> and illegitimacy<sup>111</sup> as quasi-suspect classes deserving intermediate review. A classification subject to intermediate scrutiny fails unless it is substantially related to a sufficiently important governmental interest.<sup>112</sup>

The third and highest tier of analysis, strict scrutiny, has been reserved for infringements on fundamental rights and for just a handful of suspect classifications involving race, ethnicity, and alienage.<sup>113</sup> To withstand strict scrutiny, the law or policy must be narrowly tailored to achieve a compelling governmental purpose<sup>114</sup> and use the least restrictive means.<sup>115</sup> As the lower level of analysis of rational basis review is the presumptive tier without a permitted basis for heightened review, the analysis begins there.

*a. Rational Basis Review*

Risk-needs instruments depend upon historical data extracted from group samples. Hence, risk-needs tools utilize group-based statistics, meaning that *classification*—at the heart of equal protection doctrine—is immanently embedded in contemporary risk-needs assessment. For example, a tool may, rate young, undereducated persons with a drug habit to have a higher risk of recidivism and a greater need for rehabilitative programming than people not encompassed within those groupings.

The vast majority of the classifications made by risk-needs tools are subject to rational basis review. The Supreme Court made clear that the mere recognition that a group might be stigmatized or otherwise lack equal political power is insufficient to qualify for heightened review.<sup>116</sup> To this end, courts have ruled that rational basis review is sufficient to analyze classifications based on age,<sup>117</sup> economic

---

108. *McLaughlin v. Florida*, 379 U.S. 184, 191 (1964).

109. *City of Cleburne*, 473 U.S. at 440.

110. *Craig v. Boren*, 429 U.S. 190, 197 (1976).

111. *Clark v. Jeter*, 486 U.S. 456, 461 (1988).

112. *Miss. Univ. for Women v. Hogan*, 458 U.S. 718, 724 (1982).

113. *Graham v. Richardson*, 403 U.S. 365, 372 (1971).

114. *Perry Educ. Ass’n. v. Perry Local Educators’ Ass’n*, 460 U.S. 37, 45 (1983).

115. *Bernal v. Fainter*, 467 U.S. 216, 219 (1984).

116. William N. Eskridge Jr., *Is Political Powerless a Requirement for Heightened Equal Protection Scrutiny?*, 50 WASHBURN L.J. 1, 24 (2010).

117. *Nev. Dep’t of Human Res. v. Hibbs*, 538 U.S. 721, 735 (2003); *Kimel v. Fla. Bd. of Regents*, 528 U.S. 62, 83 (2000).

status,<sup>118</sup> personality type,<sup>119</sup> mental illness,<sup>120</sup> mental disability,<sup>121</sup> and physical disability,<sup>122</sup> and applies to policies that differentiate in the treatment of alcoholics<sup>123</sup> and drug users.<sup>124</sup> Despite appealing arguments otherwise, socioeconomic class is not accorded any special status in equal protection law.<sup>125</sup>

Importantly, the rational basis test is quite deferential to government officials. To survive rational basis review a law or policy must have a legitimate purpose and be rationally related to that purpose. The government is not required to prove to the court the correctness of its judgment.<sup>126</sup> Rather, the Supreme Court affirmed that challengers must convince the court that the “facts on which the classification is apparently based could not reasonably be conceived to be true by the governmental decisionmaker.”<sup>127</sup> Even if the claimant provides evidence that the government’s judgment was mistaken, she will not prevail if the issue remains debatable in the sense that officials relied on other evidence that is at least reasonable.<sup>128</sup> Further, a court should not inquire into the correctness of the theoretical reasons for making classification distinctions as officials can still make reasonable judgments for “practical considerations based on experience.”<sup>129</sup>

In theorizing a rational basis review, one might first try to identify the likely purposes that criminal justice officials may specify for implementing risk-needs tools. The pragmatic and direct aims are to inform individual decisions concerning bail, sentencing, prison assignment, programming needs, and parole, to name just a few. The more relevant purposes for equal protection analysis, however, would be more theoretical and abstract, such as public safety, prison security, and rehabilitation. For rational basis review, the purpose merely needs to be a legitimate one. Courts have consistently and forthrightly accepted these goals as legitimate in a variety of criminal justice circumstances. In the pretrial context, the Supreme Court, reflecting on its precedence regarding classifications of pre-adjudication detainees, stated that “[a]mong the legitimate objectives recognized by the Supreme Court are ensuring a detainee’s presence at trial and maintaining safety,

---

118. *Thompson v. Gibson*, 289 F.3d 1218, 1220, 1223–23 (10th Cir. 2002) (poverty not a suspect class); *Harrison v. Bent Cnty. Corr. Facility*, 24 Fed. App’x 965, 967 (10th Cir. 2001) (indigency not a suspect class).

119. *Restucci v. Clarke*, 669 F. Supp. 2d 150, 158 (D. Mass. 2009).

120. *Heller v. Doe*, 509 U.S. 312, 321 (1993).

121. *City of Cleburne v. Cleburne Living Ctr.*, 473 U.S. 432, 442 (1985).

122. *Hibbs*, 538 U.S. at 735.

123. *Mitchell v. Comm’r of the Soc. Sec. Admin.*, 182 F.3d 272, 274 (4th Cir. 1999); *Gazette v. City of Pontiac*, 41 F.3d 1061, 1067 (6th Cir. 1994).

124. *New York City Transit Auth. v. Beazer*, 440 U.S. 568, 593–94 (1979).

125. Mario L. Barnes & Erwin Chemerinsky, *The Disparate Treatment of Race and Class in Constitutional Jurisprudence*, 72 LAW & CONTEMP. PROBS. 109 (2009).

126. *Minnesota v. Clover Leaf Creamery Co.*, 449 U.S. 456, 464 (1981).

127. *Vance v. Bradley*, 440 U.S. 93, 111 (1979).

128. *Clover Leaf Creamery*, 449 U.S. at 464.

129. *Ry. Express Agency, Inc. v. New York*, 336 U.S. 106, 110 (1949).

internal order, and security within the institution.”<sup>130</sup> A lower court authorized classification judgments in post-conviction placement and programming decisions as “there is a legitimate governmental interest to have inmates placed in [community corrections] facilities appropriate for their needs and concomitant with the public right to safety.”<sup>131</sup> Indeed, courts have routinely accepted that criminal justice officials can readily justify the higher standard of having a compelling interest in such expansive concepts of public safety,<sup>132</sup> hindering flight,<sup>133</sup> preventing crime,<sup>134</sup> and rehabilitation.<sup>135</sup> Thus, to the extent the government is convincingly able to couch its argument in terms of any one or more of these goals, the legitimate purpose portion of the test will be met. Considering that the laws and policies at issue here apply in the criminal justice system where crime control, public safety, and institutional security are core objectives, this burden of establishing a legitimate interest ought to be relatively easy to meet in most cases, except in situations where officials are relying upon truly arbitrary rationales.

Still, assuming a legitimate state interest exists, the next step is to determine whether risk-needs tools, including the factors and resulting classifications they inevitably create, are rationally related to one of the foregoing legitimate interests. From available case law, only one opinion appears to have directly addressed the use of a risk-based instrument in the context of an equal protection challenge. In the 2013 case of *People v. Osman*, the defendant argued that scoring him with the sexual recidivism risk tool Static-99 was unconstitutional.<sup>136</sup> One of the variables that Static-99 utilizes is having cohabited with an intimate partner.<sup>137</sup> A negative response is adjudged at higher risk than a positive one.<sup>138</sup> The court determined that such a distinction between groups—cohabiting v. non-cohabiting—did not implicate any protected group, such that rational basis review was applicable.<sup>139</sup>

---

130. *Collazo-Leon v. U.S. Bureau of Prisons*, 51 F.3d 315, 318 (1st Cir. 1995) (citing *Bell v. Wolfish*, 441 U.S. 520, 540 (1979) (recognizing officials possess legitimate goal of providing safe and orderly environment for inmates pretrial)).

131. *Tyler v. Pa. Bd. of Prob. & Parole*, No. 449, 2010 Pa. Commw. Unpub. LEXIS 811, at \*12–13 (Dec. 6, 2010); *see also Barge v. Pa. Bd. of Prob. & Parole*, 39 A.3d 530, 540 (Pa. Commw. Ct. 2012). (providing paroled inmates “with the proper post-incarceration treatment and surroundings is rationally related to rehabilitation and deterrence”).

132. *United States v. Chappell*, 691 F.3d 388, 399 (4th Cir. 2012); *Kachalsky v. Cacace*, 817 F. Supp. 2d 235, 269 (S.D.N.Y. 2011); *May v. Hunter*, 451 F. Supp. 2d 1084, 1088 (C.D. Cal. 2006).

133. *United States v. Salerno*, 481 U.S. 739, 754 (1987).

134. *Id.* at 749; *Schall v. Martin*, 467 U.S. 253, 264 (1984); *Bateman v. Perdue*, 881 F. Supp. 2d 709, 716 (E.D.N.C. 2012).

135. *United States v. Heath*, 419 F.3d 1312, 1316 n.2 (11th Cir. 2005); *Salerno v. Corzine*, No. 06-3547, 2013 U.S. Dist. LEXIS 141261, at \*34 (D.N.J. Oct. 1, 2013); *Smith v. Nish*, No. 3:CV-06-2291, 2007 U.S. Dist. LEXIS 37870, at \*9 (M.D. Pa. May 24, 2007).

136. *People v. Osman*, No. H037818, 2013 Cal. App. Unpub. LEXIS 2487, at \*3–4 (Cal. Ct. App. Apr. 8, 2013).

137. *Id.* at \*3.

138. *Id.*

139. *Id.* at \*14–15.

The court then summarily upheld the use of the risk tool in defendant's sentencing.<sup>140</sup> The court explained that studies had shown cohabitation experience negatively predicted sexual recidivism and, consequently, employing that factor was rationally related to legitimate interests in predicting the potential for recidivism and protecting the public.<sup>141</sup>

Despite the dearth of case opinions directly on point, other decisions in the area of criminal justice support the idea that authorities may easily link decisions they commonly make to legitimate interests, including when they are based on unprotected demographic and personal characteristics. A few examples may suffice. Regarding bail decisions that discern based on wealth-related circumstances, a scholar cites several cases to explain the reasonable assertion that "[t]he extremely permissive rational basis standard applicable to wealth discrimination would likely doom an equal protection challenge, as the bail system, for all its faults, is not wholly irrational."<sup>142</sup> Courts, in a variety of situations, have upheld classifications based on drug use, holding that the behavior is related to safety risk<sup>143</sup> and the likelihood of reoffending,<sup>144</sup> and that persons with a history of drugs require special supervision in treatment.<sup>145</sup> Similarly, a state court denied an equal protection claim of a burglary defendant who argued he was given a longer sentence than others guilty of the same offense because of his narcotics addiction; the court ruled the distinction was valid as the state had a compelling interest in providing long-term drug treatment as experience had shown that addiction and crime are correlated.<sup>146</sup> In another case example, a judge upheld under an equal protection challenge a policy that required consideration of prior drug use in decisions on prison transfers, as drug history was considered rationally related to proper institutional placement.<sup>147</sup>

Judges have found, as well, that prison officials possess proper reasons under rational basis review to distinguish violent offenders. In one case, the court concluded the prisoners "fail[ed] to establish that either their placement in the class of 'violent' offenders, their treatment within the class of violent offenders, or the difference in treatment of violent and non-violent offenders, is irrational or

---

140. *Id.* at \*15.

141. *Id.* at \*15.

142. Samuel R. Wiseman, *Pretrial Detention and the Right to Be Monitored*, 123 *YALE L.J.* 1344, 1394 n.228 (2014).

143. *New York City Transit Auth. v. Beazer*, 440 U.S. 568, 592 (1979) (finding no equal protection violation in banning methadone users from employment).

144. *In re Mabie*, 159 Cal. App. 3d 301, 308 (1984) (finding compelling interest in treating addiction to prevent drug-related crime).

145. *Beazer*, 440 U.S. at 588 n.32; *In re Lopez*, 181 Cal. App. 3d 836, 840 (1986) (without addiction cure, defendant's chance of recidivism is substantial).

146. *In re Werden*, 76 Cal. App. 3d 79, 83 (1977).

147. *Marshall v. Reno*, 915 F. Supp. 426, 432 (D.D.C. 1996).

arbitrary and not in furtherance of a legitimate governmental interest.”<sup>148</sup> In another case, the court denied an equal protection claim where the rational basis for parole authorities to separate out violent offenders was “self-evident: preventing the early release of potentially violent inmates who may pose a greater danger to the safety of others.”<sup>149</sup>

Further, though there is no evident case law directly on point, there likely is even less concern from an equal protection standpoint of the likelihood a court would rule unconstitutional the use of factors that adjudge procriminal attitudes. A person’s mindset towards antisocial causes seems reasonably relevant to a host of criminal justice outcomes, such as judgments about the individual’s culpability, likely future behavior, and amenability to supervision and treatment.

In sum, excluding for now those factors that may be subject to heightened review, it appears feasible that officials will be able to justify the use of risk-needs instruments in decisionmaking as a general rule and the vast majority of the factors within them will survive equal protection scrutiny. Several other scholars also appear to assume that risk-needs tools likely can withstand constitutional challenge (as long as race/ethnicity, and perhaps gender, are not express factors), though they generally do not undertake a comprehensive equal protection inquiry.<sup>150</sup> One scholar, however, contests this view.<sup>151</sup>

In a recent article, Sonja Starr remonstrates the vision of evidence-based sentencing practices as hardly progressive, contending current methods of risk assessment are unconstitutional when they incorporate any variables implicating race, gender, or socioeconomic status.<sup>152</sup> As for socioeconomic-related considerations, she maintains that such factors as employment, education, income, and reliance upon governmental assistance are constitutionally suspect, with her rationales interweaving equal protection and due process law.<sup>153</sup> The creative claim offered is that while the Supreme Court has not definitively found wealth to be a suspect class, the Court’s previous decisions on the matter are not as relevant to judgments regarding the use of socioeconomic status in a criminal justice context: “The treatment of indigent criminal defendants has for more than a half-century been a central focus of the Supreme Court’s criminal procedure jurisprudence. Indeed, the Court has often used very strong language concerning

---

148. *Riddle v. Mondragon*, 83 F.3d 1197, 1207 (10th Cir. 1996). Curiously, the opinion peremptorily declares the state articulated why the classifications were reasonably related to legitimate penological interests and the court declared it could think of others, yet none are listed in the opinion. *Id.*

149. *Graziano v. Pataki*, 689 F.3d 110, 117 (2d Cir. 2012).

150. Tonry, *supra* note 14, at 169 (opining that equal protection is unlikely to “impede the use of particular factors in prediction instruments” as the Court’s “jurisprudence is largely toothless as far as criminal justice system decision making is concerned”); Eaglin, *supra* note 12, at 216 (positing race/gender potentially unconstitutional factors in risk assessment); Skeem & Monahan, *supra* note 9, at 38 (generally assuming that all factors are acceptable risk factors except race).

151. Starr, *supra* note 72, at 805.

152. *Id.*

153. *Id.* at 830–36.

the importance of *eradicating* wealth-related disparities in criminal justice.”<sup>154</sup> In support thereof, the author cites two high court cases: *Griffin v. Illinois*, in which the Supreme Court struck down a requirement that convicted defendants pay court costs to receive a trial transcript, a document statutorily required to be submitted in order to appeal,<sup>155</sup> and *Bearden v. Georgia*, wherein the Court concluded that automatically revoking probation for a defendant’s inability to pay a fine was unconstitutional.<sup>156</sup> Starr points to rather broad language in these opinions to support her assertion that the Court’s intention has been to entirely “eradicat[e] wealth-related disparities” across criminal justice decisions.<sup>157</sup> In *Griffin*, the Court referred to states being prohibited in criminal trials from discriminating on the basis of poverty, just as they cannot discriminate on account of religion, race, or color.<sup>158</sup> In *Bearden*, the Court ruminated on the unfairness of punishing a person for his poverty.<sup>159</sup>

However, these two decisions do not appear adequate to sustain a broader claim that socioeconomic status can virtually never be included in a classification-oriented decision in criminal justice. A blanket prohibition on the use of wealth, much less on religion or race, would vitiate the carefully crafted three-tiered tests and otherwise thoroughly undermine the need for any equal protection analysis. Further, *Bearden* itself has been read in a far more limited manner than suggested. A few courts have rightly interpreted *Bearden* as only applying to classifications of indigency versus nonindigency as a dichotomous grouping.<sup>160</sup> Notably, the economic status-related variables in risk-needs tools do not pursue such a bifurcated structure; instead, such measures attempt to provide information about economic needs for which services can be tailored or which may correlate to failure in the community. In other cases, judges clarified that the *Bearden* ruling merely meant that probation cannot be revoked *solely* because of inability to pay.<sup>161</sup> This assessment is reasonable considering language from the *Bearden* opinion itself:

We have already indicated that a sentencing court can consider a defendant’s employment history and financial resources in setting an initial punishment. Such considerations are a necessary part of evaluating the entire background of the defendant in order to tailor an appropriate sentence for the defendant and crime. But it must be remembered that the State is seeking here to use as the

---

154. *Id.* at 830 (emphasis added).

155. *Griffin v. Illinois*, 351 U.S. 12, 18–19 (1956).

156. *Bearden v. Georgia*, 461 U.S. 660, 672–73 (1983).

157. Starr, *supra* note 72, at 830.

158. *Griffin*, 351 U.S. at 17.

159. 461 U.S. at 671.

160. *United States v. Prezioso*, 989 F.2d 52, 54 (1st Cir. 1993); *Sichenzia v. Supreme Court, Suffolk Cnty.*, No. CV-89-4348, 1990 U.S. Dist. LEXIS 1582, at \*12 (E.D.N.Y. 1990).

161. *E.g.*, *United States v. Flowers*, 946 F. Supp. 2d 1295, 1300 (M.D. Ala. 2013) (“[R]elative wealth and poverty will inevitably have some effect on the administration of justice.”); *Pedreira v. Warden*, No. 04-204-B-W, 2006 U.S. Dist. LEXIS 61718, \*13, \*18 (D. Me. 2006); *State v. Johnson*, 315 P.3d 1090, 1099 (Wash. 2014).

*sole* justification for imprisonment the poverty of a probationer who, by assumption, has demonstrated sufficient bona fide efforts to find a job and pay the fine and whom the State initially thought it unnecessary to imprison.<sup>162</sup>

Hence, even in *Bearden* the Court accepted that a sentencer could properly rely upon wealth-related information in considering punishment.

The *Griffin* ruling was also more limited than suggested. The Court later framed *Griffin* (and other relevant precedents) with the requisite circumstances that led to overturning policies requiring a fee from those unable to pay:

The individuals, or groups of individuals, who constituted the class discriminated against in our prior cases shared two distinguishing characteristics: because of their impecunty they were completely unable to pay for some desired benefit, and as a consequence, they sustained an absolute deprivation of a meaningful opportunity to enjoy that benefit.<sup>163</sup>

The *Griffin* ruling concerning wealth, therefore, required indigency *plus* a complete deprivation of a right. As a consequence, lower courts in the context of criminal justice decisions have since held that wealth classifications do not qualify for heightened review<sup>164</sup> and indigency is not itself a suspect class.<sup>165</sup> Indeed, wealth-related factors are generally considered relevant to the risk of recidivism across situations. For example, it has been opined that “[i]ncome level is not an inherently invidious basis for classification, and it is hardly irrational to conclude that a parolee without a lawful source of income is likely to return to crime to make ends meet.”<sup>166</sup> In the end, it is unlikely that equal protection law is a sufficient

---

162. 461 U.S. at 671 (emphasis added).

163. *San Antonio Indep. Sch. Dist. v. Rodriguez*, 411 U.S. 1, 20–21 (1973). The Court further opined the inability to conceptualize a definitive group of “poor” and the effect of the law not amounting to an absolute deprivation of a fundamental right meant no disadvantaged class existed deserving heightened review. *Id.* at 25.

164. *Thayer v. City of Worcester*, 755 F.3d 60, 76 (1st Cir. 2014) (panhandling ordinance); *Martinez v. Schriro*, 623 F.3d 731, 738, 742 (9th Cir. 2010) (regarding indigent defendants’ right to counsel in collateral proceedings, “the equal protection guarantee does not require the elimination of economic disparities”), *overturned on other grounds*, 132 S. Ct. 1309 (2011); *United States v. Myers*, 294 F.3d 203, 209 (1st Cir. 2002) (challenging right to appointed counsel in criminal defense); *Prows v. U.S. Dep’t of Justice*, No. 89-2929-LFO, 1991 WL 111459, at \*3 (D.D.C. June 13, 1991) (challenging a prison policy).

165. *Driggers v. Cruz*, 740 F.3d 333, 337 (5th Cir. 2014) (finding that indigent prisoners were not a suspect class); *Moore v. Unknown Part(y)(ies)*, No. 1:13-cv-669, 2014 U.S. Dist. LEXIS 69492, at \*25 (W.D. Mich. May 21, 2014) (same); *Posr v. Dolan*, No. 02 CV 0659(LBS), 2003 WL 22203738, at \*4 (S.D.N.Y. Sept. 23, 2003) (denying suspect class for *pro se* malicious prosecution litigant).

166. Paul J. Larkin, Jr., *Managing Prisons by the Numbers: Using the Good-Time Laws and Risk-Needs Assessments to Manage the Federal Prison Population*, 1 HARV. J. L. & PUB. POL’Y 1, 18 (2014) (Federalist ed.); *see also* *United States v. Kerr*, 686 F. Supp. 1174, 1179–80 (W.D. Pa. 1988) (indicating while *Bearden* may have implied a more sensitive review of wealth-based classifications in criminal justice, “lack of employment and of legitimately obtained financial resources does indicate that the defendant is likely to commit further crimes, and the deprivation of liberty may be based upon it”).



basis to preclude socioeconomic circumstances from the assessment of risk-needs, even in the criminal justice system.<sup>167</sup>

On an entirely alternative front, a critic might argue that risk-needs tools should fail even rational basis review because of numerous empirical and methodological problems therein suggesting they may not be adequately validated from a scientific perspective and thereby cannot be sufficiently related to achieve the government's goals.<sup>168</sup> If so, then perhaps the classifications are too arbitrary to withstand equal protection. Notwithstanding, a classification does not fail rational basis review simply because it "is not made with mathematical nicety or because in practice it results in some inequality."<sup>169</sup> The Supreme Court realized that "[t]he problems of government are practical ones and may justify, if they do not require, rough accommodations—[however] illogical, it may be, and unscientific."<sup>170</sup> Thus, while the science underlying risk-needs tools has been doubted by some, and there certainly may be questions about the empirical validity of some of the factors used in them, the reality is that the tools are generally accepted by the forensic mental health community and widely depended upon by experienced criminal justice officials. Their appropriateness for the decisions they inform is at the very least still debatable enough to survive the low bar of rational basis review under equal protection analysis.

### *b. Heightened Review: Gender*

Impeaching risk-needs tools under heightened review might fare better. Legal opinions differ as to whether the use of gender in risk-needs tools could survive intermediate scrutiny.<sup>171</sup> A few commentators simply assume that gender would constitute a constitutionally acceptable risk factor as a general rule.<sup>172</sup> Contrarians,

---

167. *United States v. Burgum*, 633 F.3d 810, 815 (9th Cir. 2011) (citing *Bearden* as permitting consideration of financial status in sentencing); Dawinder S. Sidhu, *Moneyball Sentencing* 37–38 (Univ. of N.M. Sch. of Law Research Paper No. 2014-26), available at [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2463876](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2463876).

168. See generally Melissa Hamilton, *Adventures in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, ARIZ. ST. L.J. (forthcoming 2015) [hereinafter Hamilton, *Adventures*], available at [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2416918](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2416918) (enumerating empirical issues with popular risk assessment instruments for violent and sexual recidivism, namely Static-99 and VRAG); Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270 (2013) (discussing logical and methodological limitations with risk tools); Melissa Hamilton, *Public Safety, Individual Liberty, and Suspect Science: Future Dangerousness Assessments and Sex Offender Laws*, 83 TEMP. L. REV. 697, 720–735 (2011) [hereinafter Hamilton, *Dangerousness*] (reviewing scientific flaws and adversarial bias in sexual recidivism risk assessment tools).

169. *Heller v. Doe*, 509 U.S. 312, 321 (1993) (internal quotes omitted).

170. *Id.* at 321 (quoting *Metropolis Theatre Co. v. Chicago*, 228 U.S. 61, 69–70 (1913)).

171. J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. REV. 1329, 1381 (2011) (surmising gender would survive equal protection analysis in risk assessments if used together with other factors).

172. Skeem & Monahan, *supra* note 9, at 38.

though, contend that considering gender in risk assessment practices could likely be judged unconstitutional.<sup>173</sup>

The doubting arguments often cite the decision of *United States v. Maples*, decided by the Fourth Circuit in 1974, in which the court “deem[ed] the factor of sex an impermissible one to justify a disparity in sentences.”<sup>174</sup> Importantly, the *Maples* court in the very same breath qualified its holding: “absent any proof that rehabilitation or deterrence are more easily accomplished in the case of females rather than males.”<sup>175</sup> The conditional is significant here as substantial disparities by gender typically exist in recidivism rates<sup>176</sup> and rehabilitation potential.<sup>177</sup> A recent study of a large cohort of prisoners in Florida, for instance, found a gendered difference in the impact of imprisonment as compared to a community sanction to recidivism.<sup>178</sup> The results suggested that imprisonment had a greater deterrence effect for women.<sup>179</sup> A meta-analysis involving multiple studies supported gendered differences, too, with researchers concluding that a longer sentence was a negative predictor of violent recidivism for male offenders but a positive predictor for women.<sup>180</sup>

Statistical correlations between gender and prison behavior, risk of recidivism, and rehabilitation potential should be sufficient to qualify as a substantial relationship to the important government interests of institutional security, prevention of crime, public safety, and programming. Admittedly, the Supreme Court in a decades-old case implicating gender discrimination rejected as insufficient the state’s statistical argument that a proportionate difference between sexes in offense rates justified disparate treatment. In *Craig v. Boren*, the state rationalized a law permitting women at a lower age than men to purchase beer, arguing the available

---

173. Starr, *supra* note 72, at 824 n.82; Eaglin, *supra* note 12, at 216. See also Carissa Byrne Hessick & F. Andrew Hessick, *Recognizing Constitutional Rights at Sentencing*, 99 CAL. L. REV. 47, 55 (2011) (noting gender an impermissible consideration in sentencing).

174. *United States v. Maples*, 501 F.2d 985, 987 (4th Cir. 1974).

175. *Id.*

176. E.g., MATTHEW R. DUROSE ET AL., DEP’T OF JUSTICE, *RECIDIVISM OF PRISONERS RELEASED IN 30 STATES IN 2005* 3 tbl.2 (2014); PATRICK A. LANGAN & DAVID J. LEVIN, DEP’T OF JUSTICE, *RECIDIVISM OF PRISONERS RELEASED IN 1994* 7 tbl.8 (2002); Jennifer E. Cobbina, et al., *Men, Women, and Postrelease Offending: An Examination of the Nature of the Link Between Relational Ties and Recidivism*, 58 CRIME & DELINQ. 331, 338 tbl.1 (2012); Hessick & Hessick, *supra* note 173, at 82 n.189 (citing studies);

177. E.g., Kelley Blanchette & Kelly N. Taylor, *Reintegration of Female Offenders: Perspectives on “What Works,”* CORRECTIONS TODAY, Dec. 2009, at 61, 62; Solveig Spjeldnes & Sara Goodkind, *Gender Differences and Offender Reentry: A Review of the Literature*, 48 J. OFFENDER REHABILITATION 314 (2009); Kirk Heilbrun et al., *How “Specific” are Gender-Specific Rehabilitation Needs?: An Empirical Analysis*, 35 CRIM. JUST. & BEHAV. 1382 (2008); Bernadette M.M. Pelissier et al., *Gender Differences in Outcomes from Prison-based Residential Treatment*, 24 J. SUBSTANCE ABUSE TREATMENT 149, 149 (2003).

178. Daniel P. Mears, et al., *Gender Differences in the Effects of Prison on Recidivism*, 40 J. CRIM. JUST. 370 (2012).

179. *Id.* at 376–77.

180. Rachael E. Collins, *The Effect of Gender on Violent and Nonviolent Recidivism: A Meta-Analysis*, 38 J. CRIM. JUST. 675, 681 (2010).

data indicated young men were far more likely to be arrested for drunk driving than young women.<sup>181</sup> The Court rejected such argument. The repudiation was not due to the statistical data being uninformative; instead, the justices simply concluded that the data were a poor fit to the state's purpose of traffic safety.<sup>182</sup> Evidence that 2.00% of young males were arrested for drunk driving (compared to 0.18% of young women) in the jurisdiction was seen as too meager a number to countenance using males as a proxy for drunk driving.<sup>183</sup> Moreover, fitness was further weakened whereby the legislation at issue prohibited the sale—but not the drinking—of beer, such that the relationship to preventing drunk driving became more attenuated.<sup>184</sup>

One commentator who maintains that using gender in risk assessments is unconstitutional conceptualizes *Craig* as standing for the propositions that equal protection “prohibit[s] . . . inferring an individual tendency from group statistics,” precludes gender-based statistical generalizations, and requires individualistic assessments.<sup>185</sup> Those abstractions seem problematic. The Supreme Court on many occasions has affirmatively approved the use of group-based statistics in decisions involving individuals. For example, the Court upheld a law criminalizing statutory rape for males only, based on broad sex-based generalizations: “Because virtually all of the significant harmful and inescapably identifiable consequences of teenage pregnancy fall on the young female, a legislature acts well within its authority when it elects to punish only the participant who, by nature, suffers few of the consequences of his conduct.”<sup>186</sup> According to the majority, the gendered classification thus was not invidious as it realistically acknowledged the sexes are not similarly situated in all circumstances.<sup>187</sup> In another case, the Court approved differential treatment between male and female naval officers whereby women were permitted a longer time for promotion as the policy “reflects, not archaic and overbroad generalizations, but, instead, the demonstrable fact that male and female line officers in the Navy are not similarly situated with respect to opportunities for professional service.”<sup>188</sup>

The Court in a later case took the opportunity to reflect upon the reasons it had permitted gendered classifications (or not) in various scenarios, summarizing that when males and females are not similarly situated because of proportionate differences in experiences or opportunities, disparities may be appropriate. Instead, “gender-based classifications . . . based solely on administrative conve-

---

181. 429 U.S. 190, 199–201 (1976).

182. *Id.* at 204.

183. *Id.* at 201–02.

184. *Id.* at 204.

185. Starr, *supra* note 72, at 825–28.

186. *Michael M. v. Superior Court*, 450 U.S. 464, 473 (1981).

187. *Id.* at 469.

188. *Schlesinger v. Ballard*, 419 U.S. 498, 508 (1975) (emphasis omitted).

nience and outworn clichés [which reflect] ‘archaic and overbroad generalizations’” will be prohibited.<sup>189</sup>

A lower court has helpfully encapsulated the high court’s case law in gender-based classifications as not requiring any mechanical test, pointing out “at least four particular matters must be explored and weighed: (1) aggregate impact on class; (2) demeaning generalizations; (3) stereotyped assumptions; and (4) flawed use of statistics.”<sup>190</sup> In this regard, *Craig* is appropriately couched as being much more about the flawed use of statistics plus a weak correlation to the government’s stated interest.<sup>191</sup> In terms of stereotypes, the Supreme Court has defined a stereotype as “a frame of mind resulting from irrational or uncritical analysis.”<sup>192</sup> Thus, the Court upheld a law that gave a monetary preference to women because of its recognition that, on average, females tended to earn less than males, and that such recognition was thereby not considered a stereotype.<sup>193</sup> These decisions approving sex-based distinctions are by their nature using group-based averages to justify the disparate treatment of protected groups, despite the likelihood that many individuals within the groups may not comport with the assumed differences. Empirically-validated and statistically-based differences in risk and needs simply do not constitute demeaning generalizations, stereotyped assumptions, or outworn clichés the justices decry when striking sex-based classifications.

The idea that empirical variations between genders supported by group level studies continue to represent proper statistics to be considered in equal protection analysis in considering if gender is substantially related to the government’s goal is further bolstered by Supreme Court case law in the area of the death penalty. In *Roper v. Simmons*,<sup>194</sup> the Supreme Court blatantly engaged group-based statistics to label an entire group, and in the process vitiated individualization. The classification in *Roper* was not gendered and it was not an equal protection case, but the reasoning is still relevant as it involved capital punishment, a legal decision uniquely individualized in its inquiry. The *Roper* court drew upon generalized statistical studies to label juveniles as lacking maturity, acting irresponsibly, behaving recklessly, being susceptible to peer pressure, and bearing an unformed character.<sup>195</sup> These broad characteristics convinced the *Roper* majority to reject the idea that a factfinder should investigate whether these traits were exhibited at the individual level and, instead, ruled that these group-based observations required the justices to consider juveniles on the whole less culpable than adults and, consequently, undeserving of the death penalty in any case.<sup>196</sup> Indeed, the dissent

---

189. *Parham v. Hughes*, 441 U.S. 347, 355 (1979) (citing *Schlesinger*, 419 U.S. at 508).

190. *Mfrs. Hanover Trust Co. v. United States*, 775 F.2d 459, 465 (2d Cir. 1985).

191. *Id.* at 467.

192. *Tuan Anh Nguyen v. INS*, 533 U.S. 53, 68 (2001).

193. *Califano v. Webster*, 430 U.S. 313, 317–18 (1977).

194. 543 U.S. 551 (2005).

195. *Id.* at 569–70.

196. *Id.* at 571.

criticized the result as using differences between juveniles and adults in the aggregate, despite the probability any such presumptions likely unbecoming many individual cases.<sup>197</sup>

As a result, the Supreme Court has not banned the use of group-based statistics in equal protection analysis, nor has it required that the government treat each individual as a wholly unique case, even in criminal justice decisions. The key will be whether officials who desire to incorporate gender into risk-needs instruments can offer studies with sufficiently strong correlations between gender and the interest at issue, be it prison behavior, recidivism, or rehabilitation potential. Any gendered differences would mean that the sexes were not similarly situated for equal protection purposes. The second aspect of *Craig* cannot be disregarded either. The classification made was a poor fit for the government's interest in preventing drunk driving as it prohibited the purchase, not the drinking, of beer. A strong statistical fit to the government's interest has been recognized in other cases as sufficient justification, notwithstanding disadvantage to a protected group. For example, the Court agreed that, despite a disproportionate impact based on race, the use of a graded test of verbal skills in qualifying for employment was acceptable where the factfinder determined a correlation between the test and performance in training existed sufficient to validate the test's usage.<sup>198</sup>

Nonetheless, there is a wrinkle with the state of current risk-needs tools in terms of gender. The instruments—to date—typically have been normed solely on males and therefore are necessarily not validated for females.<sup>199</sup> It has been rightly contemplated that ignoring gender empirically burdens the validity of risk-needs tools for use on women, even with fourth generation instruments:

Men and women are dissimilar as groups in committing crime and rehabilitation. They offend differently in many ways and respond disparately to various forms of treatment and supervision. Yet when it comes to risk assessment officials often assume they are synonymous, perhaps because of discomfort with explicit sex-based practices. Recidivism and career criminal studies consistently show that females are less involved in criminal behavior, are less likely to commit violent crimes, and are less likely to recidivate after being placed on probation or parole. Further, since the “criminal population” is largely male, any instrument that is tested on a total correctional population will naturally misclassify females.<sup>200</sup>

Forensic risk scientists and criminal justice officials have unfortunately mostly ignored these impediments, such that risk factors and criminogenic needs common

---

197. *Id.* at 601 (O'Connor, J., dissenting).

198. *Washington v. Davis*, 426 U.S. 229, 250–52 (1976).

199. ANDREW HARRIS ET AL., *STATIC-99 CODING RULES REVISED—2003*, at 5 (2003) (describing STATIC-99); VERNON L. QUINSEY ET AL., *VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK* 248 (1998) (describing VRAG).

200. James Austin, *How Much Risk Can We Take? The Misuse of Risk Assessment in Corrections*, 70 *FED. PROBATION*, no. 2, 2006, available at <http://www.uscourts.gov/uscourts/federalcourts/pps/fedprob/2006-09/risk.html>.

to women often have been excluded, and the science of risk-needs tool for women is in its infancy at best.<sup>201</sup> This state of affairs can bankrupt the imposition of any risk-needs tool that excludes gender-based considerations on women.

Other parties acknowledge that the failure to take gender into consideration, at least when predicting recidivism risk, itself is unjust. As one observer comments, “[i]ndeed, there seems to be little disputing that males, particularly relatively young men, commit more crimes, particularly violent crimes, than females of any age. If so, it would be irrational not to take those factors into account when predicting future criminality.”<sup>202</sup>

The potential unreliability of a specific risk instrument to assess women has been recognized by courts in a few cases, though not involving equal protection challenges.<sup>203</sup> In two decisions involving evidentiary attacks to sex offender registration classifications, courts recognized the underlying issue. The lower court in one decision found the state’s sex offender review board “arbitrarily and capriciously failed to evaluate evidence of the effect of gender, both on the potency of existing risk factors in predicting reoffense, and as a risk factor in its own right.”<sup>204</sup> The other case determined that available sexual recidivism risk tools and statistics for men were inapplicable to women and thus the judge expressly considered the available statistical evidence that female sexual offenders rarely reoffend.<sup>205</sup> These opinions exemplify evidentiary issues and shed light on potential equal protection issues. Further, in the event that officials were to use an instrument normed solely on males for men in an institution and not on women, it would appear that such a classification would not violate equal protection as men and women for this purpose would not be similarly situated. The instrument, validated on men, would be inapplicable to women in this regard. Hopefully, in this event, officials would be working toward norming an instrument on women to achieve its goals with respect to that group as well.

Notably, the inclusion of gender, instead of representing a negative and discriminatory purpose, actually serves the interests of institutions and defendants. Gender remains a critical classification method in criminal justice as group-based

---

201. Hannah-Moffat, *supra* note 168, at 280 (noting feminist criminologists excoriate forensic scientists for treating females as “afterthoughts”).

202. Larkin, *supra* note 166, at 18; *see also* Christopher Slobogin, *Risk Assessment and Risk Management in Juvenile Justice*, 27 CRIM. JUST. 10, 14 (2013) [hereinafter Slobogin, *Risk*] (contending age and gender constitutionally relevant in sentencing considering both related to recidivism).

203. Karsjens v. Jesson, 6 F. Supp. 3d 958, 967–68 (D. Minn.) (noting, in case challenging female’s sex offender civil commitment programming, experts’ testimony that actuarial risk tools normed on male sex offenders are inapplicable to females); *In re Risk Level Determination of S.S.*, 726 N.W.2d 121, 123 (Minn. Ct. App. 2007) (noting expert declined to score a sexual recidivism risk tool for a female defendant as it had not been validated on women).

204. *Doe v. Sex Offender Registry Bd.*, 999 N.E.2d 478, 488 (Mass. 2013).

205. *In re Coffel*, 117 S.W.3d 116, 129 (Mo. Ct. App. 2003).

statistics show that the sexes differ in risk and needs in relevant ways.<sup>206</sup> As an illustration, relevant studies regularly show that female defendants are less likely to be violent, commit a serious crime, or play a major role in crimes involving multiple offenders, and women present a lower security risk when institutionalized.<sup>207</sup> In terms of criminogenic needs, female offenders are more likely to have been violently victimized and to suffer from medical, physical, and mental problems.<sup>208</sup> Individually and collectively, these factors are relevant to culpability, predicting in-prison behavior, post-conviction functioning, and risk of antisocial attitudes, and thus should be distinctly considered when decisions are made about women as compared to men.<sup>209</sup>

To this end, evidence-based practices have appropriately evolved beyond just a half sighted focus on risk as a unitary vision of the likelihood of reoffending. Today, risk-needs tools are used to also evaluate criminogenic needs and interventions to better reduce recidivism and promote rehabilitation. Both the National Institute of Corrections and the Crime & Justice Institute promote gender-based orientations as a component of evidence-based practices.<sup>210</sup> Overall, contemporary research reinforces the idea that there are significant differences in risk and needs between genders and, as has been examined in this sub-section, with sufficient validation, variables regarding gender properly ought to be included in risk-needs tools. Plus, their inclusion should often be upheld under even heightened review so long as government officials can provide the proper empirical support between gender and the important interest at issue.

### *c. Strict Scrutiny: Race, Alienage, and Fundamental Rights*

In equal protection law, strict scrutiny applies to policies that involve classifications based on race/ethnicity and alienage or infringements on fundamental rights. Equal protection analyses regarding race/ethnicity and alienage distinguish between whether the offending policy clearly discriminates on its face versus constituting a facially neutral policy that disparately impacts a protected group.

---

206. See *supra* notes 176–180 and accompanying text; *Karsjens*, 6 F. Supp. 3d at 967–68 (recognizing expert testimony that female sex offenders differ for risk and needs purposes).

207. Kristy Holtfreter & Katelyn A. Wattanaporn, *The Transition From Prison to Community Initiative: An Examination of Gender Responsiveness for Female Offender Reentry*, 41 CRIM. JUST. & BEHAV. 41, 42 (2014) (citing studies); Emily M. Wright et al., *Gender-Responsive Lessons Learned and Policy Implications for Women in Prison: A Review*, 39 CRIM. JUST. & BEHAV. 1612, 1614 (2012) (citing studies).

208. Wright et al., *supra* note 207, at 1615–16 (citing studies).

209. *Id.* at 1617 (citing studies).

210. MADELINE CARTER, NAT'L INST. OF CORR., EVIDENCE-BASED POLICY, PRACTICE, AND DECISIONMAKING: IMPLICATIONS FOR PAROLING AUTHORITIES 8 (2011), available at <http://nicic.gov/library/024198>; CRIME & JUST. INST., IMPLEMENTING EVIDENCE-BASED PRACTICE IN COMMUNITY CORRECTIONS: THE PRINCIPLES OF EFFECTIVE INTERVENTION 3 (2004), available at <http://nicic.gov/library/019342>.

The contrast between them in practice concerns whether the court must enquire about the officials' purpose. When a classification is explicit, no inquiry into the government's intent to discriminate is required.<sup>211</sup> A facially neutral law, on the other hand, warrants strict scrutiny only if the claimant can prove that the policy was motivated by a discriminatory purpose or object, or if it is unexplainable on any other grounds.<sup>212</sup> The Supreme Court instructed that the governmental purpose to be ascertained here "implies more than intent as awareness of consequences."<sup>213</sup> A violation arises only when a public official takes an action "because of, not merely 'in spite of,' its adverse effects upon an identifiable group."<sup>214</sup>

Several scholars presume that direct measures of race and ethnicity would represent unconstitutional considerations as a general rule in criminal justice decisions.<sup>215</sup> In contrast, Michael Tonry concludes that race and ethnicity likely would be upheld as constitutional if they were among a variety of other factors being considered in risk-needs tools. In his observation, the Supreme Court's constitutional law has been "toothless" with respect to criminal justice officials' use of race/ethnicity as profiling factors.<sup>216</sup> However, Tonry admits that race and ethnicity are unlikely to be explicitly incorporated in scoring tools because they are "widely regarded as unseemly."<sup>217</sup> None of the currently popular risk-needs tools explicitly utilize either within their scored variables,<sup>218</sup> which buttresses Tonry's observation.

Nonetheless, it is worth addressing whether they could do so and still pass constitutional muster because many studies show disparities by race with both recidivism<sup>219</sup>

---

211. *Hunt v. Cromartie*, 526 U.S. 541, 546 (1999).

212. *Id.* at 546 (citations omitted).

213. *Pers. Adm'r of Mass. v. Feeney*, 442 U.S. 256, 279 (1979).

214. *Id.* (citations omitted).

215. Starr, *supra* note 72, at 812; Christopher Slobogin, *Prevention as the Primary Goal of Sentencing: The Modern Case for Indeterminate Dispositions in Criminal Cases*, 48 *SAN DIEGO L. REV.* 1127, 1168 (2011) [hereinafter Slobogin, *Prevention*]; Skeem & Monahan, *supra* note 9, at 38.

216. Tonry, *supra* note 14, at 169–70 (citing *United States v. Brignoni-Ponce*, 422 U.S. 873 (1975) (considering Mexican appearance in justifying immigration stops); see also *McCleskey v. Kemp*, 481 U.S. 279 (1987) (conceding racially disproportionate use of death penalty, defendant must still show prosecutor's racially discriminatory purpose)).

217. *Id.* Other scholars concur. "Instead of engaging in ordinary constitutional analysis when defendants challenge [sentencing] factors, courts have swept constitutional concerns under the proverbial rug based on the ungrounded conclusion that the sentencing process is somehow unique and thus shielded from constitutional review." Hessick & Hessick, *supra* note 173, at 57.

218. In the risk-needs tool in the federal post-conviction system (PCRA), ethnicity is rated but not scored. Johnson et al., *supra* note 64, at 29 app.1.

219. *E.g.*, DUROSE ET AL., *supra* note 176, at 3 tbl.2; Hessick & Hessick, *supra* note 173, at 82 n.188 (citing studies); Jeffrey Lin et al., "Back-End Sentencing" and Reimprisonment: Individual, Organizational, and Community Predictors of Parole Sanctioning Decisions, 48 *CRIMINOLOGY* 759, 776 (2010); LANGAN & LEVIN, *supra* note 176, at 7 tbl.8; ALLEN J. BECK & BERNARD SHIPLEY, *RECIDIVISM OF PRISONERS RELEASED IN 1983*, at 5 tbl.7 (1989), available at <http://www.bjs.gov/content/pub/pdf/rpr83.pdf>. But see Slobogin, *Risk*, *supra* note 202, at 14 (proving racial/ethnic factors crucial to compelling interests "unlikely, given the less-than-robust correlation



and rehabilitation outcomes,<sup>220</sup> and that criminogenic needs may vary by racial/ethnic groupings.<sup>221</sup> Moreover, where differences achieve statistical significance, the inclusion of race or ethnicity even explicitly could at least conceivably allow for material improvements in the statistical models from a predictive validity perspective and, therefore, render the tools better suited to address the compelling governmental interests in public safety, institutional security, and rehabilitative success.

Professor Oleson engages an equal protection analysis using the three-factor strict scrutiny test concerning the use of race and ethnicity in risk-needs assessments. He concludes that Supreme Court precedent suggests that the explicit use of race when found to correlate with recidivism risk may survive strict scrutiny analysis, requiring that the policy at issue be narrowly tailored to achieve a compelling government purpose and use the least restrictive means.<sup>222</sup> This conviction appears to be most befitting equal protection law analysis and as applied to the facts specifically regarding risk-needs assessments in criminal justice. For one, commentators who assert that the use of race and ethnicity are unconstitutional imply that this is true that the conclusion of illegality applies automatically and across scenarios. Yet this perspective disserves the law of equal protection. Race and ethnicity are suspect classifications, to be sure, but not entirely forbidden. Racial and ethnic groupings can survive even strict scrutiny analysis if the government meets its heightened burden. If such classifications were necessarily precluded, there would be no reason to even begin to assess whether the rationale was compelling, whether the law was narrowly tailored, or if less restrictive means were available.

The seminal case of *Regents v. Bakke*<sup>223</sup> set forth the perspective that while the explicit use of race raises great suspicions, it is not entirely forbidden. There, the Court affirmatively permitted the use of race and ethnicity as one consideration among other factors in a college admissions procedure as the state met its requisite burden under equal protection.<sup>224</sup> Recently, in the context of criminal justice, the Court in *Johnson v. United States* recognized again that even the explicit use of race can survive strict scrutiny as “special circumstances . . . may justify racial classifications in some contexts.”<sup>225</sup> Indeed, citing *Johnson*, lower courts have

---

between these characteristics and risk, as well as the large number of other risk factors available to the government”).

220. E.g., John R. Gallagher, *Drug Court Graduation Rates: Implications for Policy Advocacy and Future Research*, 31 ALCOHOLISM TREATMENT Q. 241, 247 (2013); Georgia V. Spiropoulos et al., *Moderators of Correctional Treatment Success: An Exploratory Study of Racial Differences*, 58 INT’L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 835, 836–38 (2014) (citing studies).

221. See Olaoluwa Olusanya & Jeffrey M. Cancino, *Cross-Examining the Race-Neutral Frameworks of Prisoner Re-Entry*, 20 CRITICAL CRIMINOLOGY 345, 346 (2012) (citing studies).

222. Oleson, *supra* note 171, at 1394.

223. *Regents of the Univ. of Cal. v. Bakke*, 438 U.S. 265, 320 (1978).

224. *Id.*

225. *Johnson v. California*, 543 U.S. 499, 515 (2005) (emphasis added).

found circumstances sufficient to vindicate the blatant use of race in prison cell placements.<sup>226</sup>

In any event, one can flesh out the argument that risk-needs assessments could use race/ethnicity and still pass strict scrutiny. As institutional security, public safety, and rehabilitation have qualified as compelling interests, it is appropriate to move onto the other two parts of the strict scrutiny test. Could the use of race be narrowly tailored to fulfill the goals of public safety, prison security, and rehabilitation? “Narrow tailoring does not require exhaustion of every conceivable race-neutral alternative.”<sup>227</sup> As reported earlier, many studies show that race/ethnicity is associated with reoffending rates and rehabilitation success.<sup>228</sup> As another example, a recent meta-analysis including multiple United States samples found that age, sex, and race were strongly correlated with violent recidivism in that youth, males, and non-whites were more likely to violently reoffend.<sup>229</sup>

Thus, if racial and ethnic variables significantly improved the predictive validity of risk-needs models, then including them would appear to be narrowly tailored to the government’s compelling interests. Moreover, if significant improvements in predictive ability do exist, excluding those variables undermines the state’s capability of achieving its compelling needs. Considering that one of the purposes of risk assessment is to be better able to identify, and potentially isolate, high risk or potentially violent offenders, any measure that substantially assists in that endeavor should at least not be heedlessly excluded from consideration. Notice, though, the inclusion of caveats made here. *If*, instead, scientific studies underlying a particular risk-needs tool found that race or ethnicity was not a significant correlate with the relevant outcome (recidivism, failure to appear, rehabilitation success, etc.), then developers should, practically and constitutionally, exclude it because there would be no fit with the policy’s compelling need, and certainly the use of the classification would not be narrowly tailored.

The final factor in the test includes the consideration of alternatives. There can be little doubt that criminal justice officials have over time considered and employed a plethora of options in order to achieve their compelling needs. Based on the widespread patronage of evidence-based practices across jurisdictions today, substantial evidence exists that, at least at this time, risk-needs instruments are likely the least restrictive alternative. The underlying ideology is consistent therewith. As policy analysts with profound experience in correctional interventions recognize, “[t]he risk principle states that, for the greatest impact on

---

226. *Fischer v. Ellegood*, 238 Fed. App’x 428, 434 (11th Cir. 2007); *Anderson v. Marin*, No. 1:09-cv-01547, 2012 U.S. Dist. LEXIS 181560, at \*30 (E.D. Cal. Dec. 21, 2012); *Larry v. Tilton*, No. 09-CV-0950, 2011 U.S. Dist. LEXIS 115034, at \*29 (S.D. Cal. Mar. 3, 2011).

227. *Grutter v. Bollinger*, 539 U.S. 306, 339 (2003).

228. See *supra* notes 219–220.

229. See Alex R. Piquero et al., *A Systematic Review of Age, Sex, Ethnicity, and Race as Predictors of Violent Recidivism*, 59 INT’L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 5 (2015).

recidivism, the majority of services and interventions should be directed toward higher risk individuals.”<sup>230</sup> Evidence-based practices, in attempting to reduce reliance upon incarceration and to release into the community more offenders earlier in the process, certainly appear to have a goal of infringing upon the liberty interests of fewer people and, where imprisonment is justified, to a lesser degree. The criminogenic needs portion of the third and fourth generation instruments also appear to qualify as least restrictive means in which specific rehabilitative programs are to be reserved for those with true need for them. Plus, responsivity considerations of the fourth generation further target culturally-relevant services accordingly. Hence, the later generation risk-needs tools appear to epitomize being narrowly tailored and represent the least restrictive alternative.

It is important to emphasize here that this analysis considers the use of race amongst a host of other factors within risk-needs tools. The analysis might shift if the question was whether race on its own could drive criminal justice outcomes. The argument herein draws from the recognition in the *Bakke* opinion that race could appropriately be one of many relevant factors in a decision. Still, we need not attempt such an investigation here in terms of considering whether a tool using race as the sole criterion would stand up to equal protection review. It is unlikely any tool would focus solely on race because doing so presumably would not achieve sufficient predictive ability from a statistical standpoint to justify its value. The tool would be too unitary to comply with the principles of evidence-based practices. The practice might well indicate discriminatory intent by ignoring other clearly established predictors and thus fail equal protection analysis for these reasons.

To be clear, the contention here that the direct use of race and ethnicity in risk-needs tools may be able to withstand equal protection scrutiny with strong enough empirical foundations is not meant as an unreflective recommendation *per se* that these factors must be incorporated. As will be discussed further below, the blatant use of race and ethnicity as considerations in criminal justice decisions face ethical and normative concerns.<sup>231</sup>

The final demographic variable of concern to be addressed in strict scrutiny is alienage. As the primary example, the federal Pretrial Risk Assessment includes citizenship as a predictor.<sup>232</sup> Few relevant opinions exist in available case law. In corrections, classifications involving deportable aliens have been upheld.<sup>233</sup> For instance, one court explained its rationale of treating illegal aliens disparately with

---

230. James et al., *supra* note 33, at 825.

231. The point instead is that the assumption that race and ethnicity have no legally cognizable role in risk-needs assessment is not compelled by equal protection law. A political decision to ignore them is another matter, though consequences follow. *See infra* Part III.B.

232. OFFICE OF PROBATION AND PRETRIAL SERVICES, FEDERAL PRE-TRIAL RISK ASSESSMENT INSTRUMENT: SCORING GUIDE (2013).

233. *Marshall v. Reno*, 915 F. Supp. 426, 432 (D.D.C. 1996) (upholding policy limiting community confinement options for deportable aliens).

respect to programming: “The United States may treat deportable aliens and citizens disparately. There is no primary interest in reformation of deportable persons. That’s an interest of the country to which they may be deported. Deterring further illegal reentry is a legitimate interest of the United States as well as saving expenses.”<sup>234</sup> Whether the PTRAs can withstand strict scrutiny, as well, is not so easily resolved as the instrument’s division is not set at being deportable; it rates as a positive predictor for failure any legal or illegal alien. Still, similar to the analysis with race, if this variable is significantly correlated with the interests of pretrial services in bail decisions regarding the likelihood of failure to appear, arrest, and technical violations if released, then it might survive even strict scrutiny.

Strict scrutiny outside of classifications is also reserved for policies that infringe upon fundamental rights. To date, available equal protection case law do not reveal an instance in which any actuarial tool has been excluded from informing criminal justice decisions that serve to infringe upon fundamental interests based upon arguments concerning the unfairness of including specific factors. A single case, though, is on point. In a recent case styled *People v. Osman*, the defendant argued that the actuarial risk tool Static-99 for sexual recidivism assigned points for never having lived with an intimate partner for at least two years, in violation of his First Amendment right regarding freedom of religion.<sup>235</sup> He claimed his faith as a devout follower of Islam prohibited him from living with a lover prior to marriage.<sup>236</sup> Rejecting this challenge, the court upheld the actuarial scoring as the state maintained a secular purpose of identifying a convicted sex offender’s likelihood of recidivism; further, the tool did not expressly appraise religious faith.<sup>237</sup> The Static-99 did not classify by religion on its face, yet it provides a reminder that equal protection arguments can still rely on facially neutral laws and policies.

*d. Proxies*

Disparate impact cases depend on the idea that a law or policy may be facially neutral while in effect imposing a disproportionate impact on a select group. While I have argued that risk-needs tools could survive equal protection analysis even with the most protected categories of race and ethnicity and using the stringent test of strict scrutiny (assuming the statistical footing was adequately strong), others quarrel with this notion. Some have voiced concerns that many of the factors in the instruments are merely proxies for demographic characteristics and should be eliminated on the same terms.<sup>238</sup> Scholars note that education and employment are

---

234. Ruiz-Loera v. United States, No. 00-CV-323, 2000 U.S. Dist. LEXIS 22795, at \*5 (D. Utah 2000).

235. *People v. Osman*, No. H037818, 2013 Cal. App. Unpub. LEXIS 2487, at \*4 (Cal. Ct. App. Apr. 8, 2013).

236. *Id.*

237. *Id.* at \*10–12.

238. Pari McGarraugh, Note, *Up or Out: Why “Sufficiently Reliable” Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1102 (2013) (“In order to create a risk

correlated with race and social class,<sup>239</sup> potentially even serving as statistical stand-ins for race.<sup>240</sup> Even a staunch proponent of risk-needs results in correctional decisions contends that wealth-based measures may be seen as proxies for race and, therefore, ought to be scrutinized carefully by judges as to their legitimacy.<sup>241</sup>

However, disproportionate impact, including burdening a racial minority, is not the only measure for finding unconstitutional discrimination in equal protection law.<sup>242</sup> Per the Supreme Court, “our cases have not embraced the proposition that a law or other official act, without regard to whether it reflects a racially discriminatory purpose, is unconstitutional solely because it has a racially disproportionate impact.”<sup>243</sup> The “settled rule” is that equal protection “guarantees equal laws, not equal results.”<sup>244</sup>

Thus, the Supreme Court has generally rejected proxy arguments absent proof of discriminatory intent, such as holding that a law that restricted low income housing was not regarded as intentionally targeting race, despite clear evidence of disproportionate impact on racial minorities.<sup>245</sup> The Court has not been persuaded by disproportionate results in other cases. Claimants’ “naked statistical argument” of a welfare policy’s disproportionate impact on a minority group was insufficient in itself to show the requisite racial motivation.<sup>246</sup> In another case, an employment qualification test involving verbal ability, vocabulary, and reading comprehension for police officers was upheld even though it resulted in fewer black applicants passing; the creation and implementation of the test was not deemed to exemplify a discriminatory purpose.<sup>247</sup>

Indeed, stark statistical contrasts in the impact of a policy on protected groups have not sufficed for courts to presume discriminatory intent. Thus, an employment preference given to veterans was inadequate evidence of discriminatory intent based on gender, even when ninety-eight percent of veterans were male.<sup>248</sup> In addition, a federal sentencing law requiring much longer sentences for crack cocaine defendants than powder cocaine offenders was not deemed to have a discriminatory purpose, notwithstanding evidence that ninety-four percent of

---

instrument that does not offend the Constitution, race and ethnicity, factors closely overlapping with race and ethnicity, and gender must be purged from the list of inputs.”); Hannah-Moffat, *supra* note 168, at 283; Bernard E. Harcourt, *Risk as a Proxy for Race* (Univ. of Chi. Law & Econ. Olin Working Paper No. 535, Univ. of Chi. Pub. Law Working Paper No. 323, 2010), available at [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1677654](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1677654).

239. Tonry, *supra* note 14, at 167.

240. Slobogin, *Risk*, *supra* note 202, at 14.

241. Larkin, *supra* note 166, at 18.

242. See *Washington v. Davis*, 426 U.S. 229, 242 (1976).

243. *Id.* at 239.

244. *Pers. Adm’r of Mass. v. Feeney*, 442 U.S. 256, 273 (1979).

245. *James v. Valtierra*, 402 U.S. 137, 141 (1971).

246. *Jefferson v. Hackney*, 406 U.S. 535, 548 (1972).

247. *Davis*, 426 U.S. at 245.

248. *Feeney*, 442 U.S. at 274–75.

crack offenders were black.<sup>249</sup> Other appellate courts have agreed that the disparate impact of longer crack cocaine sentences than cocaine, though a distinct proxy for race, was insufficient to constitute an equal protection violation where the evidence of racial animus or discriminatory intent by officials was, at most, contradictory, and other racially neutral reasons were provided.<sup>250</sup>

As a general rule, proxy arguments in terms of disparate impact in the context of risk-needs instruments would likely fail from an Equal Protection Clause perspective. Despite the reality that many of the variables therein disproportionately impact groups based on race, gender, and socioeconomic status, equal protection law will not itself exclude them. There is simply no evidence that the criminologists, forensic scientists, policy advocates, criminal justice officials, or politicians who have embraced evidence-based criminal justice practices in general, and risk-needs assessments in particular, did so for any reason related to a discriminatory animus of a group subject to heightened scrutiny. Certainly, the intent has been to bias high risk and violent offenders specifically, however these do not constitute protected groups, and the resulting relevant rational basis review clearly condones their disparate treatment.

## 2. Prisoners' Rights

The use of risk-needs tools to inform correctional decisions regarding security classification, institutional placement, programming, probation, parole, and supervisory conditions may implicate civil rights outside of equal protection. A criminal defendant under the supervision of criminal justice authorities, whether pretrial or post-conviction, retains his constitutional rights to the extent they “are not inconsistent with his status as a prisoner or with the legitimate penological objectives of the corrections system.”<sup>251</sup> Nevertheless, this area of constitutional law governing prisoners' rights has taken interesting and unique turns in the course of the last few decades.

For various important legal issues, the Supreme Court has adopted far more lenient standards of review for potential constitutional violations in the context of correctional practices. An exception to the leniency is sentencing, which carries its own legal structure and is addressed separately later. The decisions of correctional officials are treated differently and receive deference from the courts because, “courts are ill equipped to deal with the increasingly urgent problems of prison administration and reform.”<sup>252</sup> The Court recognized that the penal system offers a

---

249. *United States v. Blewett*, 719 F.3d 482, 487 (6th Cir. 2013), *overruled by en banc court on other grounds*, 746 F.3d 647 (6th Cir. 2013).

250. *United States v. Moore*, 644 F.3d 553, 558 (7th Cir. 2011); *United States v. Clary*, 34 F.3d 709, 713–14 (8th Cir. 1994); *United States v. Singleterry*, 29 F.3d 733, 741 (1st Cir. 1994).

251. *Pell v. Procunier*, 417 U.S. 817, 822 (1974).

252. *Procunier v. Martinez*, 416 U.S. 396, 405 (1974).

distinctly unique background in which officials are attempting to manage in a uniquely dangerous environment.<sup>253</sup>

Running a prison is an inordinately difficult undertaking that requires expertise, planning, and the commitment of resources, all of which are peculiarly within the province of the legislative and executive branches of government. Prison administration is, moreover, “a task that has been committed to the responsibility of those branches, and separation of powers concerns counsel a policy of judicial restraint.”<sup>254</sup>

As a result, judges are reluctant to intervene in issues of correctional and supervision practices.<sup>255</sup> Thus, judgments regarding prison operation and security “are peculiarly within the province and professional expertise of corrections officials, and, in the absence of substantial evidence in the record to indicate that the officials have exaggerated their response to these considerations, courts should ordinarily defer to their expert judgment in such matters.”<sup>256</sup> Prisoners’ rights law is implicated in two areas, regarding decisions that infringe upon fundamental rights or trigger due process protections.

#### *a. Fundamental Rights*

Correctional subjects do not entirely lose constitutional guarantees, though the Supreme Court has reduced the standard of review for infringements upon most of those rights to a unitary and deferential test. Per the seminal case in prisoners’ rights litigation of *Turner v. Safley*, a correctional policy that otherwise trespasses upon a constitutional right is “valid if it is reasonably related to legitimate penological interests.”<sup>257</sup> The Court explained its reasoning for such a low standard in spite of transgressing a fundamental right that would trigger heightened scrutiny in other areas of the law (such as equal protection): “Subjecting the day-to-day judgments of prison officials to an inflexible strict scrutiny analysis would seriously hamper their ability to anticipate security problems and to adopt innovative solutions to the intractable problems of prison administration.”<sup>258</sup>

The Court has used the *Turner v. Safley* rationale when evaluating claims by correctional subjects involving intrusions on such fundamental rights as speech,<sup>259</sup>

---

253. *Sandin v. Conner*, 515 U.S. 472, 482 (1995) (“[F]ederal courts ought to afford appropriate deference and flexibility to state officials trying to manage a volatile environment.”).

254. *Turner v. Safley*, 482 U.S. 78, 85–86 (1987); *see also Rhodes v. Chapman*, 452 U.S. 337, 352 (1981) (discouraging presumption that prison officials are insensitive to constitutional requirements).

255. *Safley*, 482 U.S. at 89; *Pell*, 417 U.S. at 827 (deferring to prison administrators’ implementation of policies to ensure order and security).

256. *Safley*, 482 U.S. at 86 (quoting *Pell*, 417 U.S. at 827).

257. *Id.* at 89.

258. *Id.*

259. *Beard v. Banks*, 548 U.S. 521, 528 (2006).

association,<sup>260</sup> religion,<sup>261</sup> searches,<sup>262</sup> and self-incrimination.<sup>263</sup> This means that in the realm of most correctional practices, risk-needs assessments will presumably withstand constitutional muster for a host of decisions, even if the consequences otherwise breach important individual rights. At one point, the Court generally declared that the *Turner* standard of review “applies to *all* circumstances in which the needs of prison administration implicate constitutional rights.”<sup>264</sup> Still, the Court has since clarified that the deferential stance in favor of the decisions of prison officials is subject to at least one exception: equal protection analysis of explicit race-based prison cell assignments. In the 2005 case of *Johnson v. United States* addressing automatic cell assignments based solely on race and ethnicity, a majority maintained that the permissive *Turner* test was appropriate for “rights that are ‘inconsistent with proper incarceration,’” and the “right not to be discriminated against based on one’s race . . . is not a right that need necessarily be compromised for the sake of proper prison administration.”<sup>265</sup> The decision was controversial though, with a 5:3 vote (one justice not participating) and a scathing dissent that would have retained *Turner*’s presumptive deference.

Deservedly, *Johnson* has fostered confusion about other potential exceptions to the *Turner* standard. Courts are in disagreement, for instance, about whether equal protection claims in corrections law cases regarding other protected categories, such as gender or alienage, continue to be subject to the lenient *Turner* test or instead deserve protected status.<sup>266</sup> If the answer is the former, then the government’s use of gender and alienage in risk-needs tools fare even better in the prisoners’ rights area than the previous equal protection analysis requiring a heightened review suggested. Almost certainly, an argument that significant differences in recidivism risk and criminogenic needs between genders or citizenship status is at least reasonably related to governmental interests in a correctional context, per the lax *Turner* test, could succeed given statistical justification. Thus, the use of risk-needs tools in correctional decisionmaking (distinguishing race-based variables and in the context of sentencing for now) is generally free of constitutional trouble.

There is another caveat, however. Even under the lenient *Turner* standard, there may be a cognizable challenge to risk assessment with respect to pretrial defen-

---

260. *Overton v. Bazzetta*, 539 U.S. 126, 131–132 (2003).

261. *O’Lone v. Estate of Shabazz*, 482 U.S. 342, 349–53 (1987).

262. *Hudson v. Palmer*, 468 U.S. 517, 529–30 (1984).

263. *McKune v. Lile*, 536 U.S. 24, 38 (2002).

264. *Washington v. Harper*, 494 U.S. 210, 224 (1990) (emphasis added).

265. 543 U.S. 499, 510–11 (2005) (preventing racial discrimination “bolsters the legitimacy of the entire criminal justice system [because] such discrimination is ‘especially pernicious in the administration of justice[.]’ and public respect for our system of justice is undermined” when racial discrimination is permitted).

266. Grace DiLaura, “*Not Susceptible to the Logic of Turner*”: *Johnson v. California and the Future of Gender Equal Protection Claims from Prisons*, 60 UCLA L. REV. 506, 517–18 (2012) (citing cases).



dants (as compared to post-conviction) based on the government interest in rehabilitation. Clearly, one of the main values of the latest instruments is the incorporation of a focus on identifying criminogenic needs specifically in order to change them through treatment, supervision, and services.<sup>267</sup> In other words, the needs aspect of evidence-based corrections practices is focused on improving rehabilitation potential. The Court has deemed rehabilitation programming to be a legitimate penological interest for the *Turner* test<sup>268</sup> and therefore has approved the use of risk-based classifications to tailor rules for rehabilitation purposes even though they result in infringements upon personal rights.<sup>269</sup> Lower courts have given wide latitude to prison officials in crafting treatment programs to pursue rehabilitation.<sup>270</sup> The crux of the matter, though, is that the acceptance of a legitimate governmental interest in rehabilitation distinguishes between pretrial and post-conviction defendants. The Supreme Court explained that “it would hardly be appropriate for the State to undertake in the pretrial detention period programs to rehabilitate a man still clothed with a presumption of innocence.”<sup>271</sup> Consequently, rehabilitation does not qualify as a legitimate governmental interest in a pretrial context.<sup>272</sup> As a result, at least where risk-needs instruments are utilized for any pretrial decision impacting a constitutional right, even the deferential *Turner* test would not countenance reliance upon a governmental interest in reformation. Officials will face greater difficulty in explaining the connection between variables that implicate criminogenic needs and some other interest, such as security and public safety, as the evidence-based practices literature is resplendent and consistent in its direct connection between needs (rather than risk) and reformation.

This recognition, which evidently has gone unnoticed, has a significantly unfortunate consequence to one of the important goals of evidenced-based practices, which is to situate treatment and support services earlier in the process, even pre-adjudication with pretrial programming that may permit diversion from imprisonment.<sup>273</sup> In support thereof, officials with the federal Office of Probation

---

267. LESSONS FROM THE STATES, *supra* note 36, at 6.

268. *McKune*, 536 U.S. at 37.

269. *Beard v. Banks*, 548 U.S. 521, 531–32 (2006).

270. *Newman v. Beard*, 617 F.3d 775, 781 (3rd Cir. 2010); *Lindensmith v. Petschow*, No. 12-10644, 2014 U.S. Dist. LEXIS 44721, at \*9 (E.D. Mich. Jan. 9, 2014).

271. *McGinnis v. Royster*, 410 U.S. 263, 273 (1973).

272. *Houchins v. QGED*, 438 U.S. 1, 37–38 (1978) (Stevens, J., dissenting) (noting punishment, deterrence, and rehabilitation inapplicable to pretrial detainees); *McGarry v. Pallito*, 687 F.3d 505, 513 (2d Cir. 2012) (“[I]t is clearly established that a state may not ‘rehabilitate’ pretrial detainees.”); *United States v. El-Hage*, 213 F.3d 74, 81 (2d Cir. 2000) (“Where the regulation at issue imposes pretrial, rather than post-conviction, restrictions on liberty, the legitimate penological interests served must go beyond the traditional objectives of rehabilitation or punishment.”).

273. NAT’L ASSOC. OF PRETRIAL SERVS. AGENCIES, PERFORMANCE STANDARDS AND GOALS FOR PRETRIAL DIVERSION/INTERVENTION 16–19 (2008), available at [http://www.napsa.org/publications/diversion\\_intervention\\_standards\\_2008.pdf](http://www.napsa.org/publications/diversion_intervention_standards_2008.pdf).

and Pretrial Services published an article outlining efforts to focus on appropriate treatment in the community prior to trial.<sup>274</sup> But perhaps that same document suggests acceptable alternative interests in many cases: reducing the risks of arrest, violating release conditions, and failing to appear. Of course these explanations would not necessarily save the use of risk-needs for the purpose of rehabilitation of pretrial detainees remaining in confinement.

For the foregoing reasons, constitutional debates about risk-needs tools must differentiate in the application of the law most appropriate to the context. To date, the best reading of explicit Supreme Court doctrine indicates that equal protection law and its heightened review does not apply to sentencing or to correctional decision-making outside the explicit use of race-based decisions serving prison administrative purposes. With respect to the latter, the majority in *Johnson* could have made further exceptions and was likely aware of the potential ambiguity resulting therefrom, but the fact that they did not so pontificate leaves as precedent the prior, unambiguous assertion that the *Turner* standard continues to apply outside of *Johnson*'s limited application. This means that, analyzed under prisoners' rights law, the risk-needs tools, with all of the variables currently in use (as none explicitly score race/ethnicity), are subject to the low bar of *Turner* and, therefore, likely to withstand scrutiny for the reasons stated herein. At this point, assuming risk-needs assessments pass the requisite constitutional test, the next issue relates to the idea of transparency and is addressed in the context of due process.

#### *b. Due Process*

Risk-needs assessments may implicate due process protections when they result in an infringement upon an individual's liberty interest. A claimant can derive a liberty interest from the Constitution ("by reason of guarantees implicit in the word 'liberty'") or from a statute or regulation that creates a liberty expectation.<sup>275</sup> Due process law in the correctional context has differentiated between pretrial detainees and post-conviction defendants in terms of the appropriate tests as both already involve liberty restrictions, albeit at varying degrees.<sup>276</sup> For pretrial detainees, conditions of confinement and other restrictions do not implicate due process if they are reasonably related to a legitimate and nonpunitive governmental purpose.<sup>277</sup> It has been aptly noted that the substantive due process standard for assessing pretrial detainees' claims (requiring a rational relationship to a legitimate

---

274. Cadigan et al., *supra* note 51, at 3, 5.

275. *Wilkinson v. Austin*, 545 U.S. 209, 221 (2005). For example, a state-created system granting good time credit created a liberty interest. *Wolff v. McDonnell*, 418 U.S. 539, 556–58 (1974).

276. *Bistrain v. Levi*, 696 F.3d 352, 373 (3d Cir. 2012); *Surprenant v. Rivas*, 424 F.3d 5, 17 (1st Cir. 2005); *Rapier v. Harris*, 172 F.3d 999, 1003 n.2 (7th Cir. 1999); *Mitchell v. Dupnik*, 75 F.3d 517, 523 (9th Cir. 1996); *King-Fields v. Leggett*, No. ELH-11-1491, 2014 U.S. Dist. LEXIS 21205, at \*59 (D. Md. Feb. 19, 2014).

277. *Bell v. Wolfish*, 441 U.S. 520, 538–39 (1979).

governmental objective) is akin to the *Turner* test (requiring a reasonable relationship to a legitimate penological interest).<sup>278</sup> In any event, as the foregoing due process test suggests, “[n]ot every disability imposed during pretrial detention amounts to ‘punishment’ in the constitutional sense.”<sup>279</sup> For example, conditions that are reasonably related to a penal institution’s interest in maintaining jail security typically pass constitutional muster.<sup>280</sup>

The substantive due process inquiry is distinct for defendants in a post-conviction state. In this context, due process protections are required if the restriction or deprivation either (1) creates an “atypical and significant hardship” by subjecting the subject to conditions much different from those ordinarily experienced by large numbers of inmates serving their sentences in the customary fashion, or (2) inevitably affects the duration of one’s sentence.<sup>281</sup> Opinions have somewhat fleshed out this area of law in terms of what types of correctional conditions qualify (or not) for due process protections. The Court determined that the Due Process Clause does not create a liberty interest in an inmate’s classification status or eligibility for rehabilitative or educational programs, even if the result presents a grievous loss to him.<sup>282</sup> Likewise, “conditions of confinement having a substantial adverse impact on the prisoner are not alone sufficient to invoke the protections of the Due Process Clause ‘[a]s long as the conditions or degree of confinement to which the prisoner is subjected is within the sentence imposed upon him.’”<sup>283</sup> Lower court decisions similarly have recognized that prisoners do not have a constitutionally recognized liberty interest in avoiding transfer to another placement even with the new accommodation resulting in more adverse conditions of confinement,<sup>284</sup> or in their security classification or placement,<sup>285</sup> including when the assignment is based on an assessment of future security risk.<sup>286</sup>

The Supreme Court ruled specifically that there is no liberty interest for due process purposes in a transfer from low- to maximum-security prison because “[c]onfinement in any of the State’s institutions is within the normal limits or range of custody which the conviction has authorized the State to impose.”<sup>287</sup> Conversely, the Court found certain placements in institutions that would qualify as representing “atypical and significant hardship” in conditions in two distinct scenarios. Assignment to the state’s Supermax prison required due process where

---

278. Catherine V. Struve, *The Conditions of Pretrial Detention*, 161 U. PA. L. REV. 1009, 1017 (2013).

279. *Bell*, 441 U.S. at 537.

280. *Id.* at 540.

281. *Sandin v. Conner*, 515 U.S. 472, 484, 487 (1995).

282. *Moody v. Daggett*, 429 U.S. 78, 88 n.9 (1976).

283. *Vitek v. Jones*, 445 U.S. 480, 493 (1980) (citing *Montanye v. Haymes*, 427 U.S. 236, 242 (1976)).

284. *Meachum v. Fano*, 427 U.S. 215, 225 (1976).

285. See *Hewitt v. Helms*, 459 U.S. 460, 468 (1983); *Meachum*, 427 U.S. at 228.

286. *Pacheco v. Ward*, No. 98-1104, 1999 U.S. App. LEXIS 7245, at \*4-5 (10th Cir. Apr. 14, 1999).

287. *Meachum*, 427 U.S. at 225.

Supermax was the state's most restrictive institution, inmates were held in isolated and extremely controlled conditions for indefinite periods, and the possibility of parole was suspended.<sup>288</sup> The second circumstance entailed the involuntary transfer of a prisoner to a mental health facility where the latter necessitated far greater limitations on freedom of movement, imposed significant stigmatizing consequences, and invoked "mandatory behavioral modification systems," which, together, constituted a major change in the conditions of confinement.<sup>289</sup> This Court was particularly troubled by the stigmatizing classification as mentally ill, though it also found relevant a state law that created an expectation that a prisoner would not be transferred to a mental hospital without proper procedures.<sup>290</sup>

Case law has also developed rules about liberty interests in other correctional decisions. Regarding prisoners sentenced to a term of imprisonment, the Constitution does not itself create a protected liberty interest in a pre-term expectation of parole.<sup>291</sup> However, a state's parole law or regulations could provide such an expectation and thus trigger due process protection.<sup>292</sup> Once the system grants parole, even under the condition that the individual comply with release terms, due process protections attach to the decision to revoke parole as it qualifies as a significant change in circumstances and hardship.<sup>293</sup> The same is true for probation revocation.<sup>294</sup>

Overall, many of the decisions for which authorities may use risk assessment regarding placement, transfer, prison conditions, and rehabilitation programming will qualify as reasonably related to legitimate and nonpunitive governmental purposes for pretrial subjects, and will not result in consequences that amount to an "atypical and significant hardship" for post-conviction defendants. Thus, the Due Process Clause will often not apply.

Nevertheless, as the foregoing case law review indicates, there will be times when substantive due process is triggered. Assuming a cognizable liberty interest is established and the requirement of due process invoked, the next question is what procedures are necessary to satisfy the infringement. No singular standard has emerged as there can be no one-size-fits-all procedural methods. Rather, the determination depends on the significance of the infringement, the risk of erroneous judgment, and the burdens to the state of substitute safeguards.<sup>295</sup>

---

288. *Wilkinson v. Austin*, 545 U.S. 209, 214 (2005).

289. *Vitek v. Jones*, 445 U.S. 480, 492–493 (1980).

290. *Id.* at 494; *see also* *Toevs v. Reid*, No. 06-cv-01620, 2010 U.S. Dist. LEXIS 115696, at \*16 (D. Colo. Oct. 28, 2010) (finding long-term administrative segregation program with behavioral modifications constitutes atypical and significant restraint).

291. *Greenholtz v. Inmates of Neb. Penal & Corr. Complex*, 442 U.S. 1, 10–11 (1979).

292. *Id.* at 12.

293. *Morrissey v. Brewer*, 408 U.S. 471, 482 (1972).

294. *Gagnon v. Scarpelli*, 411 U.S. 778, 782 (1973).

295. *Mathews v. Eldridge*, 424 U.S. 319, 335 (1976).

In the context of correctional decisions that rise to the level of requiring due process, the procedural requisites at times are rather minimal. In the case of finding a liberty interest in being free of assignment to a Supermax prison, the Supreme Court found acceptable policies whereby prison officials provide the inmate a brief summary of the factual basis underlying the placement decision and a “fair opportunity for rebuttal.”<sup>296</sup> These procedures were seen as commensurate “safeguards against the inmate’s being mistaken for another or singled out for insufficient reason.”<sup>297</sup> Regarding the case involving transfer to a mental hospital, the Court found adequate procedures requiring notice, time to prepare arguments, a hearing at which the inmate can present evidence and witnesses and cross-examine state witnesses, an independent decisionmaker, and a written statement of the evidence and rationale supporting a decision to transfer.<sup>298</sup>

If a state establishes an expectation of parole, the procedure approved for a parole decision included the parole board’s review of the inmate’s record and an informal interview permitting the inmate to offer letters and statements; procedural niceties not required were a formal hearing or a specification of the information in the file that led to denial.<sup>299</sup> In contrast, the minimum requirements of due process for revocation of probation or parole are far more expansive and include (a) written notice of the claimed violations; (b) disclosure of evidence against him; (c) opportunity to be heard in person and to present witnesses and documentary evidence; (d) the right to confront and cross-examine adverse witnesses (unless the hearing officer specifically finds good cause for not allowing confrontation); (e) a “neutral and detached” arbiter; and (f) a written statement by the factfinder as to the evidence relied on and reasons for revoking probation or parole.<sup>300</sup>

The procedural question at issue here is the scope of access the institution must afford to the individual’s risk-needs assessment. One might rightly ponder a defendant desiring any one or more of the following: the risk-needs outcome; scoring sheets; an accounting of the information and sources thereof the rater referenced; the instrument’s user guides and manuals; the original research the developer undertook in creating the tool; validation studies; or any other data on the tool’s predictive ability. Of course, the answer will vary depending on the breadth and extent of one’s procedural rights as just outlined. When the decision is the denial of parole where a state has created a liberty interest, the minimal procedure there did not even require a statement of information relied upon, so the prisoner likely has little right to his risk-based materials. The other types of decisions implicate greater disclosures of information and rights to challenge. Thus, in the context of placement in Supermax, the statement of facts might need

---

296. *Wilkinson v. Austin*, 545 U.S. 209, 225–26 (2005).

297. *Id.* at 226. The state’s procedure also required multiple levels of review. *Id.*

298. *Vitek v. Jones*, 445 U.S. 480, 494–95 (1980).

299. *Greenholtz v. Inmates of Neb. Penal & Corr. Complex*, 442 U.S. 1, 15 (1979).

300. *Gagnon v. Scarpelli*, 411 U.S. 778, 786 (1973); *Morrissey v. Brewer*, 408 U.S. 471, 489 (1972).

to incorporate at least the risk-needs instrument results and a “fair opportunity for rebuttal” may require more detail about scoring and the data depended upon for the particular defendant’s assessment.

With liberty infringing circumstances actuating greater procedural protections, such as the transfer to a mental hospital or probation/parole revocation, in which the government must outline the evidence upon which it relied and permit cross-examination, more disclosure is presumably necessary for procedural due process to the extent a risk-needs tool was important to the decision. Again, more than the final scores or ranking would seemingly be required. The information and sources for the data on which the assessor depended would be useful in affording the defendant a fair opportunity to challenge any factual errors. Arguably, the person(s) who conducted the risk assessment ought to be made available and the defendant given an opportunity to cross-examine in order to challenge erroneous scoring and the evaluator’s qualifications. The disclosure of supplemental materials may also be procedurally necessary to permit the defendant the ability to challenge the appropriateness from a scientific perspective of using the specific tool itself or at least to argue to the decisionmaker why so much emphasis should (or should not) be placed on the results.<sup>301</sup>

On the other hand, a court may well determine that some of the foregoing procedural niceties would improperly turn the proceedings into overly adversarial and lengthy exercises that are too burdensome from an administrative perspective. Whereas the employment of risk-needs tools in criminal justice decisions is relatively recent and legal practitioners generally have only a nascent familiarity with them such that few challenges exist to date, case law has not yet developed with respect to these procedural due process queries. It is beyond the scope of this paper to refine possible approaches, but it appeared to be befitting at least to introduce these issues for perhaps the first time.

### 3. Sentencing

The law of sentencing has generally been accorded a somewhat special stature in criminal procedure in terms of the types of information that qualify as valid considerations. On the one hand, in determinations of pretrial release, it is commonplace to evaluate residency, employment, community ties, mental health status, and substance abuse as such factors are related to the risk of failing to appear for trial and rearrest.<sup>302</sup> In addition, corrections officials can cite a

---

301. Admittedly, the presence of counsel would often be necessary pragmatically to make these types of arguments considering the intellectual difficulties the risk sciences pose. An author suggests risk instruments ought to be admissible for sentencing but not for parole decisions because only the former entails procedural protections, such as a right to counsel who can examine the appropriateness of risk tool used and the outcome, opportunity to appeal, and presence of a qualified factfinder. McGarraugh, *supra* note 238, at 1109–10.

302. THOMAS H. COHEN & BRIAN A. REEVES, U.S. DEP’T OF JUSTICE, BUREAU OF STATISTICS SPECIAL REPORT: PRETRIAL RELEASE OF FELONY DEFENDANTS IN STATE COURTS 5 (2007), available at <http://www.bjs.gov/content/pub/pdf/prfdsc.pdf>. In the federal system, judges are statutorily required in pretrial detention decisions to consider

substantial body of empirical evidence to support the use of data about criminogenic needs, requiring much information about personal and social functioning, to rather informally assign the most appropriate programming and resources to further rehabilitation success. On the other hand, the question as to whether those same factors are appropriate considerations in the adversarial stage of sentencing, with its often myopic focus on culpability, deserves its own investigation.

Relevant legal literature discloses stark disagreement as to whether future risk may be considered at all for the specific purpose of sentencing. Legal proponents stridently champion evidence-based practices as quite suited to, and comprise best practices for, sentencing proceedings.<sup>303</sup> Policy groups are on board as well. The Vera Institute, as an example, promotes judges being routinely informed by risk-needs results in determining whether a nonprison sentence is appropriate and, if so, in considering appropriate community-based services attuned to the individual defendant's needs.<sup>304</sup> A broadly subscribed initiative known as "justice reinvestment" envisions sentencing judges habitually incorporating risk-needs information in decisionmaking about whether to imprison the defendant or choose an alternative, to divert the offender to a specialty court, or to assign appropriate supervisory conditions and services during probation.<sup>305</sup> Justice reinvestment adapts the traditional judicial role to one that is not bent just on ascribing appropriate punishment in sentencing, but instead involves judges as participants in evaluating needs and responsivity per the rehabilitative side of the evidence-based model.

Critics, however, are concerned that risk-needs tools are inherently unbecoming for sentencing purposes.<sup>306</sup> A prominent criminologist expresses caution about using actuarial risk results in the sentencing process, outlining a host of methodological, pragmatic, and evidentiary issues with them.<sup>307</sup> These include the legitimacy of classifying individuals based on group data, the tendency to conflate

---

facts about "character" including physical and mental condition, family ties, employment, financial resources, length of time in the community, and community ties. 18 U.S.C. § 3142(g) (2012).

303. E.g., John Stuart & Robert Sykora, *Minnesota's Failed Experience with Sentencing Guidelines and the Future of Evidence-Based Sentencing*, 37 WM. MITCHELL L. REV. 426, 461 (2011); Roger K. Warren, *Evidence-Based Sentencing: The Application of Principles of Evidence-Based Practice to State Sentencing Practice and Policy*, 43 U.S.F. L. REV. 585, 624 (2009); Michael A. Wolff, *Evidence-Based Judicial Discretion: Promoting Public Safety Through State Sentencing Reform*, 83 N.Y.U. L. REV. 1389, 1408 (2008).

304. VERA MEMORANDUM, *supra* note 39, at 10.

305. MARSHALL CLEMENT ET AL., COUNCIL OF STATE GOV'TS, THE NATIONAL SUMMIT ON JUSTICE REINVESTMENT AND PUBLIC SAFETY 18–19 (2011), available at [https://www.bja.gov/Publications/CSG\\_JusticeReinvestmentSummitReport.pdf](https://www.bja.gov/Publications/CSG_JusticeReinvestmentSummitReport.pdf).

306. John Monahan, *A Jurisprudence of Risk Assessment: Forecasting Harm Among Prisoners, Predators, and Patients*, 92 VA. L. REV. 391, 435 (2006); Brian Netter, *Using Groups Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program*, 97 J. CRIM. L. & CRIMINOLOGY 699, 728 (2007).

307. See generally Hannah-Moffat, *supra* note 168 (citing concerns such as offense to moral and legal norms and county-specific constitutional values, the de-individualization of punishments, lack of consideration of limitations of science of risk, and unfamiliarity with technology).

correlation with causation, the questionable operationalization of the recidivism variable, the potential for race and gender discriminatory impacts, the lack of transparency in scoring, the potential need for a higher evidentiary standard if a risk tool is used to increase a sentence, and the likelihood of transferring discretion in sentencing from judges to risk tool developers.

More often the qualms are ideological in nature, drawing on the long-standing debate about the relative roles in sentencing of retributive, deterrence, utilitarian, and rehabilitative concerns. A retributive system is backward-oriented such that future predictions are innately irrelevant. John Monahan contends that risk-needs tools are appropriate for civil commitment and sexual predator civil commitment decisions (for which he claims only variables concerning race and ethnicity should be excluded), but not for sentencing.<sup>308</sup> His reason is that theoretically the focus of sentencing should be on culpability, such that concerns of future risk, being unconnected to blameworthiness, are irrelevant.<sup>309</sup> Punishment, he argues, should not consider anything else a person is (e.g., a gender), anything else a person has (e.g., a disorder), or anything else that has been done to a person (e.g., being abused as a child).<sup>310</sup> Blame attaches to what a person has done. Past criminal behavior is the only scientifically valid risk factor for violence that unambiguously implicates blameworthiness, and therefore the only one that should enter the jurisprudential calculus in criminal sentencing.<sup>311</sup>

Similarly, Paul Robinson opines that relying on even scientifically validated risk factors for future violence which do not index blameworthiness is offensive to a system of just punishment; he posits that a person does not deserve extra punishment simply because he might be young and without a father.<sup>312</sup> A commentator likewise warns that any “marginal improvements that can be gained by adding demographic considerations must be balanced against the sizable equitable costs of imposing such a regime. There is a risk in detaching punishment from the punishable act.”<sup>313</sup> U.S. Attorney General Eric Holder recently announced his opposition to the use of static and immutable characteristics in risk assessment at sentencing, arguing that punishment should be individualized to assure equal justice and further noted “they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”<sup>314</sup> A separate U.S. Department of Justice memorandum further highlights the position that risk assessment is uniquely inappropriate for sentenc-

---

308. Monahan, *supra* note 306, at 427–428.

309. *Id.* at 427–34.

310. *Id.* at 428.

311. *Id.*

312. Paul H. Robinson, *Punishing Dangerousness: Cloaking Preventive Detention as Criminal Justice*, 114 HARV. L. REV. 1429, 1440 (2001).

313. Netter, *supra* note 306, at 728.

314. Eric Holder, Attorney General, Remarks at NACDL (Aug. 1, 2014), <http://www.justice.gov/iso/opa/ag/speeches/2014/ag-speech-140801.html>.



ing purposes as it introduces an unacceptable level of uncertainty in a system that should mete out sure, swift, and fair punishments.<sup>315</sup> A deterrence regime in a sentencing system would also find risk-needs data ill-suited, at least to the extent they are based on immutable characteristics which cannot be altered and are thus not deterrable.

In opposition, sentencing regimes adopting alternative philosophies with future orientations would find predictions palatable. A more utilitarian jurisdiction would adjudge risk-needs tools greatly attractive, perhaps even necessary, to achieve instrumental goals.<sup>316</sup> To the extent a sentencing system incorporated prison population reduction targets, it inherently seeks the ability to identify low-risk candidates for community corrections. A sentencing scheme embracing rehabilitation as a proper objective would present the most accommodating regime to risk-needs assessments.

Admittedly, no definitive answer overall can be given here about the suitability of risk-needs to sentencing from an ideological perspective as legislatures are lawfully welcome to adopt any one or more of the foregoing theories in their sentencing laws and policies. The resolution based simply on ideological grounds, thus, may vary by jurisdictional requisite.

Regardless of the jurisdiction's underlying sentencing philosophy, interested observers note that unease about the types of variables used in risk-needs tools are heightened in the context of sentencing as compared to other criminal justice decisions.<sup>317</sup> Michael Tonry concludes that factors such as race, ethnicity, religion, and gender may properly be used for decisions as to culturally-appropriate program assignments, yet be unsuitable for decisions involving punishment.<sup>318</sup> Few proponents or critics have seemed to notice one particular legal pitfall.<sup>319</sup> The use of certain demographic variables in risk-needs tools potentially violates state statutes. Sentencing laws in many states require that sentence decisions be neutral of a variety of status variables, including race, ethnicity, national origin, gender, and religion, and some preclude other characteristics which would make risk-needs assessments even more vulnerable considering the host of variables within the tools that implicate them, of social status and economic status.<sup>320</sup>

---

315. Letter from Jonathan J. Wroblewski, Director, Office of Policy and Administration, Department of Justice, to Judge Patti B. Saris, Chair, U.S. Sentencing Commission 7 (July 29, 2014), available at <http://www.ussc.gov/sites/default/files/pdf/training/annual-national-training-seminar/2014/doj-annual-report.pdf>. The Department of Justice distinguishes the use of risk-needs assessments in sentencing, for which it opposes, but lauds them for reentry purposes. *Id.* at 1–8.

316. Slobogin, *Prevention*, *supra* note 215, at 1159–60.

317. Tonry, *supra* note 14, at 171 (problematizing demographic and lifestyle choice factors “less acute in contexts other than sentencing.”).

318. *Id.*

319. *But see* Sidhu, *supra* note 167, at 28–29 (noting many risk-needs variables violate federal sentencing statutes); Hannah-Moffat, *supra* note 168, at 283 (noting same).

320. *E.g.*, ARK. CODE ANN. § 16-90-801(b)(3) (2013) (listing race, gender, social, and economic status); FLA. STAT. § 921.002(1)(a) (2013) (listing race, gender, and social and economic status); MASS. GEN. LAWS ch. 211E,

Sentencing may differ from other criminal justice decisions for another reason. A controversy continues as to whether the tests of scrutiny applied to protected groups under equal protection law are relevant in the first instance to sentencing challenges. Many commentators and judges simply assume, as an example, that race-based considerations at sentencing are absolutely prohibited—without exception.<sup>321</sup> Some case opinions have taken a broad swath, asserting a defendant's race "may play no adverse role in . . . sentencing."<sup>322</sup> The constitutional origin of such an absolute ban is unclear. Other courts convey the legal situation that is likely more accurate, reflecting the use of race in sentencing as still subject to the Equal Protection Clause whereby strict scrutiny applies.<sup>323</sup> The Tenth Circuit perhaps provides the best interpretation of the state of the law here in recognizing that strict scrutiny still applies to the use of race in sentencing, citing the Supreme Court's criminal justice decision in *Johnson* applying strict scrutiny to race-based prison cell assignments.<sup>324</sup> If the Tenth Circuit's conclusion is correct, then the analysis of the use of variables of race, ethnicity, and alienage provided previously in the equal protection analysis pertains equally to sentencing. In opposition, if those who believe that race is automatically forbidden as a sentencing consideration are right, such a legal ruling cannot be explained by equal protection law necessitating an analysis of governmental objectives and need, even in heightened scrutiny. Thus, assuming the Supreme Court at some future time were to expressly impose a sort of strict liability bar to any consideration of a protected category at sentencing, the ruling would most assuredly reflect judicially-imposed reasons of *public policy*, rather than any sort of traditional constitutional analysis.<sup>325</sup>

The nature of the legal tests for protected groups aside, few cases appear to have addressed the use of risk-needs instrument results in determining criminal punishment. An Indiana appellate court at one point ruled that reliance upon the structured professional judgment instrument's (LSI-R) results applied in the case to aggravate punishment was improper because the tool merely represented algorithmic data and constituted an exercise that failed to exemplify an appropriate substitute for an independent analysis of the facts, an exercise which sentencing

---

§ 3(e) (2012) (listing race, sex, national origin, creed, religion, and socio-economic status); MICH. COMP. LAWS § 769.34(3)(a) (2013) (listing gender, race, ethnicity, alienage, national origin, and employment); NEV. REV. STAT. § 176.0125(3)(f) (2013) (listing race, gender, and economic status); OHIO REV. CODE ANN. § 2929.11(C) (LexisNexis 2013) (listing race, ethnic background, gender, and religion); TENN. CODE ANN. § 40-35-102(4) (2013) (listing race, gender, creed, religion, national origin, and social status).

321. Sidhu, *supra* note 167, at 35 (citing *Zant v. Stephens*, 462 U.S. 862, 885 (1983)); Oleson, *supra* note 171, at 1379.

322. *United States v. Kaba*, 480 F.3d 152, 156 (2d Cir. 2007); *United States v. Leung*, 40 F.3d 577, 586 (2d Cir. 1994).

323. *Gonzales v. Cockrell*, No. MO-99-CA-072, 2002 U.S. Dist. LEXIS 28988, at \*76–77 (W.D. Tex. Dec. 19, 2002).

324. *United States v. Smart*, 518 F.3d 800, 804 n.1 (10th Cir. 2008) (opining use of race in sentencing decision would not violate equal protection if compelling reasons to justify it).

325. *United States v. Lyman*, 261 F. App'x 98, 100 (10th Cir. 2008).

decisions demand.<sup>326</sup> However, this treatment of risk-needs assessment results was shortly thereafter effectively overturned by the state's supreme court. In *Malenchik v. State*, the higher court affirmatively encouraged evidence-based practices as a whole, and as a part thereof favored the ability of sentencers to use information from risk-needs tool results in order to craft appropriate sentencing options with an eye toward fostering reformation.<sup>327</sup> In response to Malenchik's argument that it was incorrect for a sentencer to consider socioeconomic factors, which were a component of LSI-R, the higher court responded that state rules required socioeconomic information in pre-sentence reports and such facts were relevant at sentencing to understanding the likelihood of recidivism and criminogenic needs.<sup>328</sup> Thus, at least there is some precedent in favor of risk assessment in sentencing and the pertinence of socioeconomic factors therein.

Assuming that risk-needs assessments are appropriate evidence for sentencing purposes, a separate rationale may distinguish sentencing from other correctional decisions from the perspective of transparency. Defendants in general enjoy greater procedural rights at sentencing than in other correctional situations. The question posed here is what rights do sentencing defendants have in receiving evidence regarding the risk-needs tool if one was relied upon in the sentencing process? This query parallels the previous discussion in the prisoners' rights arena as to the level of access a defendant might enjoy to information about scoring, the tool's guides, validation studies, etc. Yet, in sentencing, more robust procedural rights must mean that a defendant is entitled to a greater degree of disclosure than in any prisoners' rights circumstance, including at least some information about the risk-needs component of a sentencing decision.

Several courts have assumed defendants enjoy no due process right to have access to *all* the information on which the sentencing decisionmaker based its decision.<sup>329</sup> Still, due process requires that information relied upon in sentencing be relevant, reliable, and accurate.<sup>330</sup> The Supreme Court itself ruled that a sentence formed on materially-untrue assumptions about the defendant's criminal history violates due process.<sup>331</sup> Thus, courts have found that a defendant must be given the factual information on which the sentencer relied and a meaningful opportunity to rebut it.<sup>332</sup> It is also noted that a sentence must normally be vacated

---

326. *Rhodes v. State*, 896 N.E.2d 1193, 1195 (Ind. Ct. App. 2008). The court also complained LSI-R rated factors duplicating considerations the sentencing judge would already have included in the base punishment, such as criminal history, education, and employment. *Id.*

327. 928 N.E.2d 564, 574 (Ind. 2010).

328. *Id.* at 574–75.

329. *Stewart v. Erwin*, 503 F.3d 488, 495 (6th Cir. 2007); *United States v. Curran*, 926 F.2d 59, 62 (1st Cir. 1991).

330. *Gov't of Virgin Islands v. Yarwood*, 45 V.I. 68, 77 (V.I. 2002).

331. *Townsend v. Burke*, 334 U.S. 736, 740–41 (1948).

332. *United States v. Millán-Isaac*, 749 F.3d 57, 70 (1st Cir. 2014); *Smith v. Woods*, 505 F. App'x 560, 568 (6th Cir. 2012); *United States v. Hayes*, 171 F.3d 389, 394 (6th Cir. 1999).

where the sentencing judge relies on prejudicial pre-sentence material from unidentified sources that the defendant was not given an opportunity to rebut.<sup>333</sup> More specifically as to the issues presented herein, information on which a judge relies in determining the defendant's potential for future dangerousness ought to be disclosed.<sup>334</sup> The *Malenchik* opinion, mentioned earlier, confirmed this position, pontificating a bit on the defendant's access to a risk assessment's scoring sheet completed by probation as part of its pre-sentence investigation:

A defendant is entitled to a copy of the pre-sentence report prior to his sentence being imposed . . . . Thus the defendant will be aware of any test results reported therein and may seek to diminish the weight to be given such test results by presenting contrary evidence or by challenging the administration or usefulness of the assessment in a particular case.<sup>335</sup>

As risk tools become offered in a greater number of sentencing cases in the future, assuming they pass the potential constitutional barriers discussed herein, it is likely that the number of defendants gaining access to risk assessment information and attempting to rebut the information underlying the scoring and to correct scoring, even to challenge the applicability of the tool itself as the *Malenchik* decision implies, will soar. This should be an advantage to litigants generally. Transparency is valued at sentencing and, overall, more knowledge should be publicized and shared about the advantages and deficits of risk-needs methodologies across criminal justice decisions. In sum, the evaluation of risk-needs information may be differentially oriented in sentencing as compared to other criminal justice contexts. The differing theoretical purposes of sentencing yield varying results. Retributive and deterrence orientations are less amenable to evidence-based practices while utilitarian and rehabilitative foci would embrace them. Practitioners must also be wary of whether certain factors may violate sentencing statutes. With clarity, due process concerns are heightened in sentence decisionmaking whereby more information on risk-needs assessments ought to be provided to defendants and shared among the relevant professional communities to foster understanding and further improve evidence-based practices.

### *B. Ethical and Normative Concerns*

In addition to voices claiming that certain aspects of risk-needs tools are illegal, many contend that they contain a host of factors that should be deemed unethical—regardless of their constitutionality—to use in a criminal justice context. One concern is that risk-needs tools may serve to punish normative lifestyle choices that individuals in a free society are otherwise at liberty to make, such as whether

---

333. *United States v. Huff*, 512 F.2d 66, 71 (5th Cir. 1975).

334. *United States v. Hamad*, 495 F.3d 241, 246 (6th Cir. 2007).

335. *Malenchik v. State*, 928 N.E.2d 564, 575 (Ind. 2010) (citation omitted).

to marry, pursue an education, remain employed, or purchase a home.<sup>336</sup> The ethics-based complaints most often center around the idea that immutable characteristics should be excluded, such as race, ethnicity, religion, gender, and perhaps age.<sup>337</sup> A scholar cites generalized human rights legislation that prohibits the use of age, sex, race, and disability in discriminatory ways.<sup>338</sup> Another believes that the idea of “[p]aying a penalty justified only by an immutable personal characteristic runs counter to nationwide trends in equity and imposes serious societal costs,” including detaching punishment from the culpable act, unfortunately segregating individuals within predictive groups, and suffering many false positives.<sup>339</sup>

Other commentators are likewise concerned with the idea of culpability. It may appear unethical and immoral to base decisions that impact liberty interests on immutable characteristics considering individuals bear no responsibility for them,<sup>340</sup> or on any other characteristic for which the individual has little control, such as mental or physical health status.<sup>341</sup> A quotation from the Supreme Court may support this idea, whereby equal protection law is at times concerned with a classification based on an immutable characteristic which its possessors are powerless to escape or set aside. While a classification is not *per se* invalid because it divides classes on the basis of an immutable characteristic, it is nevertheless true that such divisions are contrary to our deep belief that “legal burdens should bear some relationship to individual responsibility or wrongdoing,” and that advancement sanctioned, sponsored, or approved by the State should ideally be based on individual merit or achievement, or at the least on factors within the control of an individual.<sup>342</sup>

In contrast, scholars who favor risk-needs tools in criminal justice provide counter arguments. Despite the quotation just given, it is notable that equal protection law does not absolutely prohibit the use of protected categories if there is a legally cognizable reason to differentiate on those bases.<sup>343</sup> It may be the case, too, that the use of static factors that individuals cannot control may be justified as

---

336. Tonry, *supra* note 14, at 171.

337. Netter, *supra* note 306, at 716–17.

338. Ivan Zinger, *Actuarial Risk Assessment and Human Rights: A Commentary*, 46 CANADIAN J. CRIMINOLOGY & CRIM. JUST. 607, 611 (2004).

339. Netter, *supra* note 306, at 728.

340. Tonry, *supra* note 14, at 171.

341. CHRISTOPHER SLOBOGIN, PROVING THE UNPROVABLE: THE ROLE OF LAW, SCIENCE, AND SPECULATION IN ADJUDICATING CULPABILITY AND DANGEROUSNESS 113 (2007); Thomas Nilsson et al., *The Precarious Practice of Forensic Psychiatric Risk Assessments*, 32 INT’L J. L. & PSYCHIATRY 400, 406 (2009) (“A basic demand on just legislation is that all offenders are to be treated equally and fairly, which is hardly the case judging from the way society has singled out the category of mentally disordered subjects as especially perilous. They are supposed to be extensively scrutinised and, when there is a risk for relapse into criminality, they are handed over to an unlimited form of detention with considerably reduced individual rights.”).

342. *Regents of the Univ. of Cal. v. Bakke*, 438 U.S. 265, 360–61 (1978) (citation omitted) (quoting *Weber v. Aetna Casualty & Surety Co.*, 406 U.S. 164, 175 (1972)).

343. Larkin, *supra* note 166, at 18.

they may simply be proxies to other, more palatable risk-based characteristics. Relatedly, Christopher Slobogin explains that

risk-based dispositions are ultimately based on a prediction of what a person will do, not what he or she is. Immutable risk factors are merely *evidence* of future conduct, in the same way that various pieces of circumstantial evidence that are not blameworthy in themselves (e.g., presence near the scene of the crime, possession of a weapon) can lead to a finding of guilt.<sup>344</sup>

There is the pragmatic approach as well. For example, because being male and of a young age consistently correlate with recidivism, it might be unreasonable not to include these factors as predictor variables.<sup>345</sup>

The foregoing concerns are often not couched in legal terms *per se* but are largely political in nature in recognition of social consequences. They have been persuasive in some cases. A staunch proponent of risk-needs instruments observes that indigency seems relevant to whether, for example, a person on parole may resort to crime, but he is willing to be more politically correct: “If, however, there is too great a risk that correctional officials might use poverty as camouflage for race, then courts can carefully scrutinize use of that particular feature or eliminate it altogether without condemning risk-needs assessments in the process.”<sup>346</sup>

Several developers of risk-needs tools have succumbed to these sociopolitical concerns. The developers of an actuarial risk tool for sentencing purposes noted they intentionally excluded race and ethnicity as variables, vaguely referring to “stakeholder sensitivities.”<sup>347</sup> The developers of HCR-20 were forthright about the matter: “Some risk factors, despite showing statistical associations with violence in the population, may be considered *prima facie* objectionable to include in an assessment for the purpose of estimating violence risk. Examples include race, gender, and minority ethnic status.”<sup>348</sup> Virginia officials developed the state’s own risk instrument, in the end intentionally excluding race as a rated variable, despite its statistically significant correlation with recidivism; interestingly, their justification was based on race as a proxy for social and economic disadvantage rather than the reverse.<sup>349</sup> As another example of political correctness, the creators of the federal system’s post-conviction risk tool (PCRA) purposely removed gender from the final instrument, even though their original regression model found being

344. Slobogin, *Risk*, *supra* note 202, at 15.

345. Larkin, *supra* note 166, at 18.

346. *Id.*

347. Richard Berk & Justin Bleich, *Forecasts of Violence to Inform Sentencing Decisions*, 30 J. QUANTITATIVE CRIMINOLOGY 79, 87 (2014), available at [www-stat.wharton.upenn.edu/berkr/SentCART%20copy.pdf](http://www-stat.wharton.upenn.edu/berkr/SentCART%20copy.pdf).

348. Kevin S. Douglas et al., *Historical-Clinical-Risk Management-20, Version 3 (HCR-20<sup>V3</sup>): Development and Overview*, 13 INT’L J. FORENSIC MENTAL HEALTH 93, 96 (2014).

349. Richard P. Kern & Meredith Farrar-Owens, *Sentencing Guidelines with Integrated Offender Risk Assessment*, 16 FED. SENT’G REP. 165, 165 (2004).

female was statistically significant as a negative predictor of recidivism.<sup>350</sup> Developers who are so motivated generally have reacted by resorting to regrettably unsophisticated and unempirical methods by merely eliminating ethically questionable predictors without compensating for the lost predictive value.<sup>351</sup>

Overall, the promise of evidence-based practices envisioned by many policy groups, forensic risk investigators, criminal justice officials, and academics has been foreshadowed by equally fierce animosity by other professionals within those same fields. The censure of risk-needs instruments variously espouses constitutional challenges and moralized objections. The criticisms have appeared to convince at least some developers to simply boycott what might otherwise be significant predictors from their models to appease censors.

#### IV. THE FUTURE OF SOCIODEMOGRAPHIC FACTORS

This Article has outlined various constitutional, ethical, and normative objections to risk-needs instruments. Many have objected to the incorporation of various factors immutable in nature—thereby unchangeable—and thus deemed offensive or otherwise create apprehension when they form the basis of risk prediction in criminal justice decisions. In the sentencing regime, opponents voice philosophical opposition as representing improper considerations in sentencing regimes that ought to be focused on culpability.<sup>352</sup>

Regarding the purportedly offensive factors, race and ethnicity appear to cause the greatest unease. Bernard Harcourt is entirely against prediction models to reduce prison populations because he views risk as merely a proxy for race.<sup>353</sup> Critics also target gender, other immutable characteristics, and socioeconomic factors. Of course, the million-dollar question is what to do since evidence-based practices essentially rely upon empirical risk-needs tools? The clear alternatives are (1) to go all in, employing any empirically validated tool regardless of the factors therein; (2) to cease risk-needs assessments altogether, as Harcourt suggests; or (3) to choose something in between, such as eliminating politically offensive variables.<sup>354</sup>

The third option attracts much attention. Choosing this posited alternative of simply jettisoning disquieting factors comes with unfortunate consequences to the overall platform and aspirations of evidence-based practices. Empirical value will necessarily be compromised as the tools typically include only variables found to

---

350. Johnson et al., *supra* note 64, at 19 tbl.1. PCRA creators simply noted that subsequent analyses determined the variable involving gender did not sufficiently improve the predictive validity of the model overall. *Id.* at 22.

351. Stephen D. Gottfredson & Laura J. Moriarty, *Statistical Risk Assessment: Old Problems and New Applications*, 52 *CRIME & DELINQ.* 178, 194 (2006).

352. *See supra* notes 309–316 and accompanying text.

353. Harcourt, *supra* note 238.

354. *Id.* at 9.

be statistically significant to the risk or need of interest. Simply discarding politically sensitive variables and their proxies from risk-needs tools can critically jeopardize predictive ability.<sup>355</sup> The values of empiricism, objectivity, and transparency also depreciate when sociopolitical concerns are elevated over science. To the contrary, a commentator who contends that risk-tools instruments should exclude, for equal protection reasons, variables related to race, gender, and wealth-based factors asserts that doing so would not compromise the predictive validity of tools as a general rule. The support cited by the commentator as justification is a single study that ostensibly “suggests that demographic and socioeconomic factors could be excluded from risk prediction instruments without losing any significant predictive value.”<sup>356</sup> This conceptualization of the research overreaches for several empirical reasons.

First, the cited study investigated a subset of a database compiling information on defendants sentenced in 1980 in a few counties from a single state.<sup>357</sup> Thus, the study appears too old and too geographically limited to be generalizable. Second, the study actually did not test any risk instrument, or anything analogous to one. The research predated most second-generation risk assessment tools and all third and fourth generation tools. Instead of testing any existing tool, the researchers examined a surfeit of criteria that at the time were often used by various criminal justice officials to make unstructured judgments about risk across the areas of sentencing, probation supervision, and parole guidelines.<sup>358</sup> Thus, the study cannot stand as a representative example of any actuarial based model or structured professional judgment tool and the results cannot be generalized across past, existing, or future instruments. Third, regarding the allegation that none of the demographic or socioeconomic factors held predictive ability, the study’s findings on the full sample only showed that a few of the race-correlated status variables failed to improve predictive ability in the full sample. Several status factors uncorrelated with race had already been included in the statistical analysis and, together with other untainted factors, already had been shown to perform better than chance.<sup>359</sup> As the authors of the study themselves concluded, “dropping status factors from guidelines would do very little to reduce racial disparities in sentencing, probation supervision, and parole decisions. It might, in fact, increase them by removing criteria that make a greater number of white offenders look like bad risks” because most of the race-correlated status variables affected white

---

355. Hannah-Moffat, *supra* note 168, at 284.

356. Starr, *supra* note 72, at 851 (citing Joan Petersilia & Susan Turner, *Guideline-Based Justice: Prediction and Racial Minorities*, 9 CRIME & JUST. 151, 161 (1987)).

357. Petersilia & Turner, *supra* note 356, at 161.

358. *Id.* at 158 tbl.1.

359. *Id.* at 171 fig.1. The untainted status variables included high school graduate, mental illness, age, employed, and living with a spouse. *Id.* at 164–65 tbl.2 (referencing full sample of prisoners and status variables designated with \* as not correlated with race).



felons adversely.<sup>360</sup> On the whole, this study is insufficient on its own to justify broader claims and it is still the case that removing statistically significant demographic variables, particularly a large number of them, would reduce predictive ability.

Curiously, some scholars who most staunchly object to the incorporation of many of the variables in risk-needs instruments remain willing to retain criminal history.<sup>361</sup> If the argument is that gender and race, any proxies for gender and race, and/or socioeconomic factors should be excluded because it is simply unjust or politically incorrect to use them for criminal justice decisions, then the same assumption seemingly ought to apply at least as equally to criminal history. Studies consistently show that criminal history is strongly correlated with gender, race, and socioeconomic factors.<sup>362</sup> One of these authors at least attempts to explain the apparent contradiction by arguing that criminal history is distinguishable as one has personal autonomy and control over committing crimes.<sup>363</sup> This contrast is nebulous, as individuals do not entirely lack the ability to alter their sociodemographic positions. Further, the assumption that criminal history measures used in risk-needs tools only cover incidents in which the person actually committed the criminal offense scored is amiss. Many of the tools count as criminal history any evidence of offending, even without some formal confirmation such as a conviction, an arrest, or an official record of any sort; most still count juvenile crimes and offenses for which the individual was officially exonerated as well.

More ideological reasons caution against excluding variables for reasons other than empirical weakness or failure to be reasonably related to governmental interests. Simply excising significant factors begins to grievously undermine other core foundations of evidence-based practices. Recall that the advancements most favored in the third and fourth generation of instruments were the incorporation of dynamic factors, the philosophy that criminogenic needs should be addressed to reduce recidivism and that attention should be focused on responsiveness to culturally-relevant services. The factors that are the subject of criticism are generally the same factors that are highly relevant to needs and responsiveness, thereby to decisions fostering successful rehabilitative programming. Importantly, recent studies typically show that culturally-sensitive considerations of race, ethnicity, gender, and other immutable factors are necessary to improve rehabilitation results.<sup>364</sup> Eliminating these factors from the risk-needs assessment necessarily undermines gathering information appropriate to connecting needs and responsiv-

---

360. *Id.* at 166.

361. Sidhu, *supra* note 167, at 70; Starr, *supra* note 72, at 872.

362. Hannah-Moffat, *supra* note 168, at 283 (citing studies regarding gender, race, and socioeconomic factors); King, *supra* note 31, at 547 (citing the influence of race).

363. Sidhu, *supra* note 167, at 66.

364. See generally D.A. ANDREWS & JAMES BONTA, *THE PSYCHOLOGY OF CRIMINAL CONDUCT* (5th ed. 2010) (listing studies).

ity to supervision, programs, and services.

As for the philosophical grievances with respect to sentencing specifically, prohibiting risk-needs consideration at all in sentencing decisions poses another assault on a fundamental goal of evidence-based practices, which is the incorporation of judges at sentencing into the broader enterprise. Justice reinvestment includes molding the judicial role at sentencing into one in which risk-needs data can inform judges when considering whether to imprison, the desirable length of sentence, the appropriateness of alternative sanctions, and the choice of probation conditions and service needs. Either stance against evidence-based practices, whether because of a philosophical focus on culpability or because of legal and ethical concerns about certain variables therein, severs or curtails this crucial component of the evidence-based model reliant upon judicial involvement and participation.

Admittedly, elsewhere I have argued in opposition to the incautious dependence upon actuarial risk assessments for criminal justice decisions because of empirical concerns about reliability and validity.<sup>365</sup> My editorials included tribulations about certain risk instruments having been normed on foreign samples yet indiscriminately scored on domestic offenders, high rates of false positives, exaggerations of predictive validity measures, evidence of adversarial bias in scoring, the lack of standardization in sufficiently training raters, and the inherent inability of group-based statistics to permit individualized predictions of risk.<sup>366</sup> These concerns remain, and because they persist, it is even more deeply troubling that heedlessly removing statistically significant factors further renders the instruments increasingly less reliable and valid from a statistical perspective. Here, though, the contention is that as long as embedding risk-needs instrument results to inform criminal justice decisions continues to be performed in practice, then at least it makes sense to permit officials to rely upon the best science available, instead of destabilizing the very foundations upon which evidence-based practices emerged.<sup>367</sup>

In any event, a counter perspective might point out the potential unfairness where, if officials are obeying the empirically-driven dictate that a tool should not be used to rate a person or group for whom it was not validated, many individuals or groups cannot then be so assessed. In other words, they may be treated differently, which raises legal suspicions and undermines uniformity and consistency. These concerns are real, but should not be dispositive. First, it must be recognized that we have no national uniformity in criminal justice practices in the

---

365. See generally Hamilton, *Adventures*, *supra* note 168; Hamilton, *Dangerousness*, *supra* note 168.

366. See Hamilton, *Adventures*, *supra* note 168; Hamilton, *Dangerousness*, *supra* note 168.

367. I am little concerned that using immutable traits, even gender and race, in risk-needs will be viewed by many as evidence of animus or indicative of any derogatory discriminatory purpose. There are simply no signs that evidence-based practices embody any nefarious intent toward any group except for those offenders for whom empirically validated instruments using a variety of variables rate as high risk. Simply ignoring an abundance of evidence that differences exist in risk, needs, and responsiveness, even with race, gender, and socioeconomic status, sacrifices empiricism for political correctness.

first instance. A defendant may be rated on a risk-needs tool in one state but not in another simply due to variations in state practices. An inmate in one jail may be subject to a risk-needs analysis whereas an inmate in another jail even in the same city may not. No equal protection problem arises, though, as these inmates are not similarly situated.<sup>368</sup> Second, even if groups within the same institutional placements are differentially rated on risk measures because of validation concerns, there is also no evident issue to the extent they also are then not similarly situated. The benefits of evidence-based practices should not be suspended until the instruments are validated on everyone in the institution. Third, even if the institution's practice is to assess all offenders anyway, there is the potential for overrides to the extent the assessor considers the individual to differ in some risk or needs-relevant way(s) not addressed by the tool. Finally, in terms of potentially disserving groups based on immutable characteristics, an additional consequence of removing those factors is that the practice diminishes officials' ability to protect potential future victims sharing those same characteristics as studies typically indicate that offenders often commit crimes against those with similar demographic and status traits.<sup>369</sup>

Another value will be lost by abridging evidence-based practices: innovation. If we cease risk-needs assessments or abbreviate them by removing important variables to assuage political sensitivities, we lose valuable information, experience, and data that scientists could mine to greatly improve their models and use to conduct further studies in order to cross-validate the instruments on more and more groups. Advancements in empirically driven risk-needs tools are critical to criminal justice decisions. As has been recognized,

[t]he application of accurate and up-to-date information, including all known and empirically validated risk factors, thereby ensures that hearing examiners have the tools they need to arrive at individualized classification determinations. "Such determinations must be grounded in a corpus of objective facts and data, necessarily dynamic and evolving to revise collective understanding of the risk that various individuals pose to the public."<sup>370</sup>

The deeper the knowledge researchers are able to accumulate and study, the greater progress evidence-based practices can achieve. Evidence-driven decisions are seen to hold the key to reducing reliance on over-incarceration, targeting services to offenders who most need them, and reducing recidivism. If risk-needs tools are censored or if constitutionally or ethically suspect variables are excised therefrom, it is likely that fact-finders would consider risk and the factors of race,

---

368. *Beaulieu v. Ludeman*, No. 07-1535, 2008 U.S. Dist. LEXIS 119324, at \*37 (D. Minn. Feb. 8, 2008) ("Detainees at one facility or unit are not considered to be 'similarly-situated' to detainees at other facilities or units for Equal Protection purposes.")

369. Berk & Bleich, *supra* note 347, at 13.

370. *Doe v. Sex Offender Registry Bd.*, 999 N.E.2d 478, 489 (Mass. 2013).

gender, and socioeconomic status in criminal justice decisions informally anyway, rendering their use less reliable, transparent, and consistent.<sup>371</sup>

## V. CONCLUSIONS

It may be somewhat true that assessing risk is akin to predicting the winner in a horse race.<sup>372</sup> Still, criminal justice officials rightly seek out scientifically validated methods to enhance risk prediction capabilities and gauge criminogenic needs. The review herein of the evolution of risk-needs instruments highlighted interdisciplinary advancements among numerous private and public industries. These endeavors have, at the same time, spawned controversies. Evidence-based practices in criminal justice represent, contradictorily, either a panacea of best practices or a harbinger of unfair and unconstitutional biases. Risk-needs instruments incorporate a host of variables that are scientifically shown to be statistically significant, yet many of them also inflame certain political sensitivities. The utility of risk-needs instruments also varies depending upon the type of criminal justice decision and whether its preferred philosophy underlying it is retributive, deterring, incapacitative, or rehabilitative in nature. Legal scholars, forensic professionals, and policy analysts continue to struggle with these paradoxes.

This article reviewed these constitutional and moral quandaries for the use of risk-needs assessment across multiple criminal justice decisions. Certainly, hard choices must be made. But this state of affairs is not a new predicament in criminal justice. Trying to make amends for a history of discrimination can lead officials to sacrifice when making decisions to improve public safety in order to appease stakeholder and public sensitivities. Further, policymakers continue to debate the most appropriate philosophical orientation to employ. In the end, after critically analyzing the various legal and political arguments, I conclude that modern risk-needs methodologies—assuming empirical validation and statistical significance—need not for constitutional or moral reasons be forsaken or truncated. The country holds compelling reasons to innovate to curb its record incarceration rate, offer appropriate rehabilitation, and improve institutional safety, to the mutual benefit of all.

---

371. Richard S. Frase, *Recurring Policy Issues of Guidelines (and non-Guidelines) Sentencing: Risk Assessments, Criminal History Enhancements, and the Enforcement of Release Conditions*, 26 FED. SENT'G REP. 145, 151 (2014).

372. QUINSEY ET AL., *supra* note 199, at 36.

## APPENDIX A: POPULAR RISK ASSESSMENT TOOLS

1. *Second Generation Tools*

<b>Instrument</b>	<b>Factors Rated</b>
Violent Risk Appraisal Guide (VRAG) <sup>373</sup>	• Nonviolent criminal history score
	• Failure on prior conditional release
	• Age
	• Marital status
	• Lived with both biological parents to age 16
	• Elementary school maladjustment
	• Alcohol problems
	• Victim Injury
	• Female victim
	• Personality disorder
	• Schizophrenia
Static-99 <sup>374</sup>	• Number of prior sex offense charges
	• Prior convictions for a non-contact sex offense
	• Convictions for an index non-sexual violence
	• Convictions for non-sexual violence before index
	• Number of prior sentencing dates
	• Age
	• Lived with intimate partner for 2 years
	• Nonfamilial victims
	• Stranger victims
	• Male victims

---

373. *Id.* at 237–238.

374. ANDREW HARRIS ET AL., *STATIC-99 CODING RULES REVISED—2003* (2003).

<b>Instrument</b>	<b>Factors Rated</b>
Federal Pretrial Risk Assessment (PTRA) <sup>375</sup>	• Number of prior convictions
	• Number of prior failure to appears
	• Any pending cases
	• Current offense type
	• Current offense class
	• Age
	• Citizenship
	• Highest education
	• Employment status
	• Own residence
	• Current drug problems

## 2. Third Generation Tools

<b>Instrument</b>	<b>Factors Rated</b>
HCR-20 <sup>376</sup>	Historical
	• Previous violence
	• Prior supervision failure
	• Young age at first violent incident
	• Relationship instability
	• Employment problems
	• Early maladjustment
	• Substance use problems
	• Major mental illness
	• Psychopathy
	• Personality disorder

375. OFFICE OF PROBATION AND PRETRIAL SERVICES, FEDERAL PRE-TRIAL RISK ASSESSMENT INSTRUMENT: SCORING GUIDE (2013).

376. Kevin S. Douglas & Christopher D. Webster, *The HCR-20 Violence Risk Assessment Scheme: Concurrent Validity in a Sample of Incarcerated Offenders*, 26 CRIM. JUST. & BEHAV. 3, 8 (1999).

Instrument	Factors Rated
HCR-20 (cont.)	Clinical
	• Lack of insight
	• Negative attitudes
	• Active symptoms of major mental illness
	• Impulsivity
	• Treatment nonresponse
	Risk Management
	• Plans lack feasibility
	• Exposure to destabilizers
	• Lack of personal support
	• Noncompliance with remediation
• Stress	
LSI-R <sup>377</sup>	Criminal History
	• Prior adult convictions
	• Number of current offenses
	• Arrested before age 16
	• Prior incarceration
	• Escape history
	• Punished for institutional misconduct
	• Community supervision violation
	• History of violence
	Education/Employment
	• Employment history
	• Educational attainment
	• School suspensions
	• Participation in school activities
	• Peer interactions
	• Interactions with authorities

---

377. NEW SOUTH WALES DEPT. OF CORRECTIVE SERV., LSI-R TRAINING MANUAL (2002).

Instrument	Factors Rated	
LSI-R (cont.)	Financial	
	• Financial problems	
	• Reliance on social assistance	
	Family/Marital	
	• Dissatisfaction with marital situation	
	• Interaction with parents	
	• Criminal family	
	Accommodations	
	• Residential stability	
	• High crime neighborhood	
	Leisure/Recreation	
	• Participation in organized activity	
	• Appropriate use of time	
	Companions	
	• Socially isolated	
	• Criminal acquaintances	
	Alcohol/Drugs	
	• Alcohol problems	
	• Drug problems	
	• Alcohol/drugs contributed to law violations	
	• Family alcohol/drug use	
	Emotional/Personal	
	• Distress	
	• Psychosis	
	• Mental health treatment	
	• Prior psychological assessment	
	Attitudes/Orientation	
	• Procriminal attitudes	
• Prosocial orientation		
• Attitude toward sentence		
• Attitude toward supervision		



### 3. Fourth Generation Tools

Instrument	Factors Rated
Federal Post Conviction Risk Assessment (PCRA) <sup>378</sup>	• Number of prior arrests
	• Prior community supervision violations
	• Institutional adjustment problems
	• History or current violent offense
	• Varied offending pattern
	• Age
	• Married
	• Highest education level
	• Employment status
	• Work history
	• Alcohol problems
	• Drug problems
	• Family problems
	• Lack of social support
• Motivated to change	
COMPAS <sup>379</sup>	• Criminal involvement
	• History of noncompliance
	• History of violence
	• Current violence
	• Criminal associates
	• Substance abuse
	• Financial problems
	• Vocational or educational
	• Family criminality
	• Social environment
	• Leisure

378. Johnson et al., *supra* note 64.

379. THOMAS BLOMBERG ET AL., VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT 14–15 (2010).

<b>Instrument</b>	<b>Factors Rated</b>
COMPAS (cont.)	• Residential instability
	• Social isolation
	• Criminal attitudes
	• Criminal personality

Running head: RACE, RISK & RECIDIVISM

**Risk, Race, & Recidivism:  
Predictive Bias and Disparate Impact**

Jennifer Skeem

University of California, Berkeley

[jenskeem@berkeley.edu](mailto:jenskeem@berkeley.edu)

and Christopher T. Lowenkamp

Administrative Office, U.S. Courts

[christopher\\_lowenkamp@ao.uscourts.gov](mailto:christopher_lowenkamp@ao.uscourts.gov)

Corresponding author: Jennifer Skeem, University of California, Berkeley, 120 Haviland Hall  
#7400, Berkeley, CA 94720-7400

\* The views expressed in this article are those of the authors alone and do not reflect the official position of the Administrative Office of the U.S. Courts. Lowenkamp specifically advises against using the PCRA to inform front-end sentencing decisions or back-end decisions about release without first conducting research on its use in these contexts, given that the PCRA was not designed for those purposes.

## Abstract

One way to unwind mass incarceration without compromising public safety is to use risk assessment instruments in sentencing and corrections. Although these instruments figure prominently in current reforms, critics argue that benefits in crime control will be offset by an adverse effect on racial minorities. Based on a sample of 34,794 federal offenders, we examine the relationships among race, risk assessment (the Post Conviction Risk Assessment [PCRA]), and future arrest. First, application of well-established principles of psychological science revealed little evidence of test bias for the PCRA—the instrument strongly predicts arrest for both Black and White offenders and a given score has essentially the same meaning—i.e., same probability of recidivism—across groups. Second, Black offenders obtain higher average PCRA scores than White offenders ( $d = 0.34$ ; 13.5% non-overlap in groups' scores), so some applications could create disparate impact. Third, most (66%) of the racial difference in PCRA scores is attributable to criminal history—which is already embedded in sentencing guidelines. Finally, criminal history is *not* a proxy for race, but instead mediates the relationship between race and future arrest. Data are more helpful than rhetoric, if the goal is to improve practice at this opportune moment in history.

**Key words:** risk assessment, race, test bias, disparities, sentencing

## Risk, Race, & Recidivism: Predictive Bias and Disparate Impact

Over recent years, increased awareness of the economic and human toll of mass incarceration in the U.S. has launched a reform movement in sentencing and corrections (see Lawrence, 2013). This remarkably bipartisan movement (Arnold & Arnold, 2015) is shifting public discourse about criminal justice “away from the question of how best to punish, to how best to achieve long-term public safety” (Subramanian, Moreno, & Broomhead, 2014, p. 2).

One way to begin unwinding mass incarceration without compromising public safety is to use risk assessment instruments in sentencing and corrections. These research-based instruments estimate an offender’s likelihood of re-offending, based on various risk factors (e.g., young age, prior arrests)—and they figure prominently in current reforms (Monahan & Skeem, in press). Across the U.S., statutes and regulations increasingly require that risk assessments inform decisions about the imprisonment of higher-risk offenders, the (supervised) release of lower-risk offenders, and the prioritization of treatment services to reduce offenders’ risk (National Conference of State Legislators, 2015; see also American Law Institute, 2014). By implementing risk assessment at sentencing, Virginia diverted 25% of nonviolent offenders from prison without raising the crime rate (Kleiman, Ostrom & Cheesman, 2007).

Despite such promising results, controversy has begun to swirl around the use of risk assessment in sentencing. The principal concern is that benefits in crime control will be offset by costs in social justice—i.e., a disparate and adverse effect on racial minorities and the poor. Although race is omitted from these instruments, critics assert that risk factors that are sometimes included (e.g., marital history, employment status) are “proxies” for minority race and poverty (Harcourt, 2014; Starr, 2014; Silver & Miller, 2002). In the view of Former Attorney General Eric Holder (2014), risk assessment

“may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society. Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant’s history of criminal conduct. They should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place.”

These concerns are legitimate and important—but untested. In fact, Holder specifically urged that this issue be studied. The main issue is whether the use of risk assessment in sentencing affects racial disparities in imprisonment, given that young black men are six times more likely to be imprisoned than young white men (Carson, 2015). Risk assessment could *exacerbate* racial disparities, as Holder speculates. But risk assessment could instead have *no effect* on—or even *reduce* disparities—as others have predicted (Hoge, 2002: see also Gottfredson & Gottfredson, 1988).

It must be understood that concerns about racial disparities are more-or-less applicable to all uses of risk assessment in sentencing and corrections. Although criticism focuses on the use of risk assessment to inform *front-end* sentences that judges impose, the same concerns are applicable to *back-end* sentencing decisions about release from incarceration (earned release, parole, etc.). Regardless of the decision’s timing (front- or back-end) or type (to release lower-risk offenders or to detain higher-risk offenders)—there could be a net effect of risk assessment on racial disparities in incarceration. Even the well-established use of risk assessment to inform resource allocation in corrections (see Elek, Warren, & Casey, 2015) can invoke concern. If higher-risk offenders are subject to more intensive community supervision and risk reduction

services—and service refusal violates the terms of release—they are more subject to social control than their lower-risk counterparts.

Does risk assessment exacerbate, mitigate, or have no effect on racial disparities? The answer to this question probably depends on factors that include the instrument chosen. Sensationalistic headlines aside, “risk assessment” is not reducible to “race assessment” (Sentencing Project, 2015). Validated risk assessment instruments differ in their purpose and in the risk factors they include (Monahan & Skeem, in press)—and little is known about their association with race.

In the present study, we use a cohort of federal supervisees to empirically test the nature and strength of relationships among race, risk assessment scores, and recidivism. Because existing disparities in punishment “primarily affect black Americans” (Tonry, 2012, p. 54), we focus on Black and White offenders. Our goal is to inform debate and provide guidance for instrument selection and refinement. To contextualize this study, we first highlight where risk assessment fits in corrections and sentencing, and then unpack controversy about particular types of risk factors.

### **Risk Assessment in (Community) Corrections**

Risk assessment has been used to inform correctional decisions for nearly a century (Administrative Office of the U.S. Courts, 2011). Early instruments were designed to achieve efficient prediction; they generally involved scoring a set of risk markers, weighting them by predictive strength, and combining them into a risk score that could be used to rationalize the use of supervision resources (e.g., assigning higher risk offenders to more intensive community supervision). Later instruments have often been infused with the concept of risk reduction: They include variable risk factors as “needs” to be addressed in supervision and treatment and are

meant to scaffold principles of evidence-based correctional services. These principles specify who should be treated (those at relatively high risk of recidivism, given the “risk” principle) and what should be treated (variable risk factors for crime, given the “need” principle).

Decades ago, Gottfredson et al. (1994; Gottfredson & Jarjoura, 1996) noted the potentially discriminatory effects of risk assessment in justice settings (see Petersilia & Turner, 1987) and illustrated how to remove “invidious predictors.” Since then, little concern has been expressed about such correctional applications. In fact, risk assessment plays a central role in The Sentencing Reform and Corrections Act of 2015, a bill before congress that requires that risk assessments be conducted to assign federal inmates to appropriate recidivism reduction programs (e.g., work and education programs, drug rehabilitation). Inmates who comply with these programs can earn early release (for up to 25% of their remaining sentence).

### **Where Risk Assessment Fits in Punishment Theory**

Front-end applications of risk assessment attract the greatest controversy. Since the mid-1970’s, sentencing in the U.S. has largely been a backward-looking exercise focused on an offender’s moral blameworthiness for the conviction offense, in keeping with retributive theories of punishment (Monahan & Skeem, in press). Over recent years, sentencing reform has reflected a resurgence of interest in incorporating forward-looking assessments of an offender’s risk of future crime, in keeping with utilitarian or crime control theories of punishment.

Currently, risk assessment is considered—and in our view *should* be considered—within bounds set by moral concerns about culpability (Monahan & Skeem 2014). This is consistent with the leading model of criminal punishment (Frase, 2004)—a hybrid of retributive and utilitarian theories called “limiting retributivism” (Morris, 1974). As operationalized in the Model Penal Code (American Law Institute, 2014), sentencing takes place “within a range of



severity proportionate to the gravity of offenses, [and] the blameworthiness of offenders.” Within this range, a sentence is chosen to promote “offender rehabilitation [and] incapacitation of dangerous offenders” (§1.02(2), p. 2). That is, retributive concerns set a permissible range for the sentence (e.g., 5-9 years), and risk assessment is used to select a particular sentence within that range (e.g., 8 years for high risk). Risk assessment should never be used to sentence offenders to more time than they morally deserve.

### **Controversial Risk Factors**

**Risk factors irrelevant to blameworthiness (Starr & socioeconomic factors).** The retributive task of assigning blame for past crime and the utilitarian task of assessing risk for a future crime are orthogonal—but it is easy to make category errors (Monahan & Skeem, in press). This tendency to conflate risk with blame constrains the risk factors perceived as appropriate to consider at sentencing. The least controversial variable—criminal history—relates to blame and risk in similar ways: Past involvement in crime aggravates perceived blameworthiness for a conviction offense *and* increases the likelihood of future offending. More controversial variables like low educational attainment do not bear on an offender’s blameworthiness for a conviction offense (e.g., someone who did not complete high school is no more blameworthy than someone who did), but do increase the risk of recidivism.

According to Starr (2014, 2015), it is legitimate to consider an offender’s criminal history in determining a sentence—but risk assessment instruments also include such “socioeconomic” variables as marital history, employment/education, and financial background. In her view, these variables are illegitimate—*both* because they are unrelated to moral culpability *and* because they are perceived as “proxies” for poverty and minority status. In Starr’s arguments, blame eclipses risk, as a concern appropriate to consider at sentencing.

**Risk factors associated with race (Harcourt’s & criminal history).** In sharp contrast to Starr, Harcourt (2008) objects to the use of criminal history to inform sentencing, whether the vehicle is sentencing guidelines (which emphasize criminal history) or risk assessment instruments (which typically include criminal history alongside other risk factors). In Harcourt’s view (2015) “criminal history has become a proxy for race.”

Minority race and criminal history are correlated (e.g., Durose, Snyder & Cooper, 2015; Petersilia & Turner, 1987)—although the degree varies as a function of how criminal history is operationalized. For example, in a meta-analysis of 21 studies, Skeem, Edens, Camp & Colwell (2004) found negligible differences ( $d = .06$ ) between Black and White groups on a multi-item criminal history sub-scale that robustly predicts recidivism (Walters, 2012). Moving from research to practice, Frase, Roberts, Hester, & Mitchell (2015) found that sentencing guidelines vary substantially in their operationalization of criminal history. Data from four jurisdictions indicate that Black offenders obtain higher average criminal history scores than White offenders (*Mean*  $d = .24$ , *SD* = .05)—with the range of effect sizes ( $d = .19-.29$ ) suggesting about 79%-85% overlap between groups (see Cohen, 1988).<sup>i</sup>

Criminal history reflects not only the differential participation of racial groups in crime (e.g., Black people being involved in crime—particularly violent/serious crime—at a higher rate than Whites), but also the differential selection of given groups by criminal justice officials (e.g., police decisions about arrest; prosecutor decisions about charging) and by sentencing policies (e.g., minimum mandatories; Blumstein 1993; Frase, 2009; Tonry & Melewski, 2008; Ulmer, Painter-Davis & Tinik, 2014). The proportion of racial disparities in crime explained by differential participation vs. differential selection is hotly debated (see Frase 2014; McCord, Widom & Crowell, 2001), and varies as a function of crime type (e.g., violence vs. drug crimes)

and stage of justice processing (e.g., arrest vs. incarceration; Blumstein et al., 1983; Piquero, 2015).

**Risk factors that cannot be changed (Holder’s & “static” characteristics).** Starr (2015) suggests that risk factors “within the defendant’s control” may legitimately be considered in sentencing. Although she does not articulate how to distinguish risk factors that reflect life choices from those that mark hapless socioeconomic circumstance (a fraught task; see Tonry, 2014), her suggestion mirrors Holder’s (2014) view that the most objectionable risk factors for the purposes of sentencing are “static” and “immutable” characteristics (except criminal history).

Risk assessment instruments oriented toward risk reduction explicitly include variable risk factors that can be shown to change through intervention. For example, substance abuse problems and criminal thinking patterns (e.g., feeling entitled, rationalizing misbehavior) are robust risk factors that can be treated to reduce recidivism (Monahan & Skeem, 2014). Variable risk factors may be perceived as less problematic than fixed markers that cannot be changed (e.g., young age at first arrest) and variable markers that cannot be changed through intervention (e.g., young age).

**Summary.** Legal scholars who oppose the use of risk assessment at sentencing find risk factors that may be associated with race particularly objectionable when they are irrelevant to (or mitigate) an offender’s blameworthiness or cannot be changed. As is clear from this brief review, critics disagree in calling potentially race-related risk factors like criminal history “in” or “out,” for the purposes of sentencing.

### **Bringing Psychological Science to the Controversy**

**Test bias vs. disparate impact.** Data may be more helpful than rhetoric, if the goal is to improve sentencing and correctional practices at this opportune moment in history. Ample

guidance on racial fairness in assessment is available from similar efforts undertaken in more mature fields (e.g., intelligence and other cognitive tests used to inform high-stakes education and employment decisions, see Reynolds 2000; Sackett, Borneman & Connelly, 2008). There is substantial agreement on the empirical criteria that indicate when a test is biased. These criteria have been distilled in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014)—which we refer to as the “Standards.”

Given that the *raison d'etre* for risk assessment instruments is to predict recidivism, the paramount indicator of test bias is *predictive bias* (also known as “differential prediction;” Standard 3.7). On utilitarian grounds alone, any instrument used to inform sentencing must be shown to predict recidivism with similar accuracy across groups. If the instrument is unbiased, a given score will also have the same meaning regardless of group membership (e.g., an average risk score of X will relate to an average recidivism rate of Y for *both* Black and White groups). This is commonly tested by examining whether groups systematically deviate from a common regression line that relates test scores to the criterion (Cleary, 1968; see also Sackett & Bobko, 2010).

Given a pool of instruments that are free of predictive bias, however, some instruments will yield greater mean score differences between groups than others (e.g., Black people, on average, will obtain higher risk scores than Whites). These instruments are not necessarily biased: “subgroup mean differences do not in and of themselves indicate lack of fairness” (The Standards, #3.6, p. 65). The notion that mean differences are indicative of test bias is unequivocally rejected in the professional literature because group differences in scores may reflect true differences in recidivism risk, based on group variation “in experience, in

opportunity, or in interest in a particular domain” (Sacket et al., 2008, p. 222). Race reflects longstanding patterns of social and economic inequality in the U.S. (e.g., differences in social networks/resources, neighborhoods, education, employment). Although poverty and inequality do not inevitably lead to crime, they “involve circumstances that do contribute to criminal behavior” (Walker, Spohn, & DeLone, 2011, p. 99). Group differences in such circumstances can manifest as valid group differences in risk scores.

Even if mean score differences do not reflect test bias, using instruments that yield such differences to inform sentencing may create *disparate impact* (in legal terms; see *Griggs vs. Duke Power*, 1971 cf. *McClesky v. Kemp*, 1987) or inequitable social consequences (in moral terms; Reynolds & Suzuki 2012). Simply put, even if an instrument perfectly measured risk, *use* of the instrument could still be seen as unfair. As Frase (2013) observes, even when racial disparity “...results from the application of seemingly appropriate, race-neutral sentencing criteria, it is still seen by many citizens as evidence of societal and criminal justice unfairness; such negative perceptions undermine the legitimacy of criminal laws and institutions of justice, making citizens less likely to obey the law and cooperate with law enforcement” (p. 210). For such reasons, the Standards (3.6) suggest that instruments be examined to understand and (if possible) reduce group differences. If two instruments are equally valid “and impose similar costs,” the Standards (3.20) advise “selecting the test that minimizes subgroup differences.”

In our view, risk assessment instruments used at sentencing—and the risk factors they subsume—must be empirically examined for both predictive bias and disparate impact. Simply put, risk assessment must be both empirically valid and perceived as morally fair across groups.

This study is among the first to rigorously examine the relations among risk, race, and recidivism among adult offenders in the U.S. Although this issue has been studied with juvenile

offenders (e.g., Olver et al., 2009), forensic instruments designed to predict violence (e.g., Singh & Fazel, 2010), and indigenous/non-indigenous groups in other countries (e.g., Wilson & Gutierrez, 2014), our focus is on comparing Black and White offenders in the U.S. on instruments designed to predict recidivism. In a recent meta-analysis, Desmarais, Johnson, & Singh (in press) identified 53 studies of 19 risk assessment instruments used in U.S. correctional settings. Only three studies permitted comparisons of predictive accuracy by offender race—and indicated that levels of predictive utility were identical (Area Under the ROC Curve or AUCs=.69 on the “COMPAS;” Brennan et al., 2009) or highly similar (Odds Ratio or ORs=1.03 [Black] and 1.04 [White] on the Levels of Services Inventory-Revised or LSI-R; Lowenkamp & Bechtel, 2007; Kim, 2010) across groups. Formal tests of predictive bias were not reported, nor were mean score differences.

**Proxies vs. mediators.** Beyond defining bias in testable terms, science can also lend precision to discourse about—and understanding of—controversial risk factors. Risk assessment critics often use the term “proxy” to refer to some risk factors. Calling criminal history a proxy for race (Harcourt, 2015) suggests that the two variables are so highly correlated that criminal history can be used as an indirect indicator of race—to “stand in” when race is not measured directly. However, it is rarely clear that factors like criminal history are *meant* to proxy for race (i.e., to camouflage discrimination).

Progress is possible when terms like “proxy” are operationally defined. Kraemer et al. (2001) clarify how risk factors can work together to predict an outcome like recidivism. In their terminology, a proxy is a correlate of a strongly predictive risk factor that also appears to be a risk factor for the same outcome—but the only connection between the correlate and the outcome is the strong risk factor correlated with both. By their criteria, criminal history is a

proxy for race only if race “dominates” in predicting recidivism (i.e., maximum strength in predicting recidivism is achieved by race alone – not criminal history alone; not the combination of criminal history and race). This is unlikely, given that criminal history typically predicts recidivism much more strongly than race (Berk, 2009; Durose et al., 2014). In this study, we apply Kraemer et al’s (2001) criteria to determine whether criminal history is a proxy for race—or instead, possibly mediates race’s relation to recidivism (i.e., is correlated with race and explains much of the relationship between race and recidivism).

### **Present Study**

In the present study, we use a cohort of Black and White federal offenders to empirically examine the relationships among race, risk assessment, and recidivism. In the federal system, risk assessment is not used to inform front-end sentencing decisions. Instead, the Post Conviction Risk Assessment or “PCRA” (Johnson, Lowenkamp & VanBenschoten, 2011) is administered upon intake to a term of supervised release to inform decisions designed to reduce offenders’ risk—i.e., to identify *whom* to provide with the most intensive supervision and services (higher-risk offenders) and *what* to target in those services (variable risk factors). The PCRA was developed by the US Administrative Office of the Courts to improve the effectiveness and efficiency of federal community supervision—and should not be used for other sanctioning purposes unless and until it is validated for those purposes.

The PCRA is well-validated and includes major risk factors tapped by many other risk assessment instruments—including criminal history (the subject of Harcourt’s objection); education, employment, and social network problems (central to Starr’s objection); and other variable factors (e.g., substance abuse, attitudes) that have drawn less controversy. These federal data can address aims with broader implications:

1. To what extent is the instrument—and the risk factors it includes—free of *predictive bias*? We hypothesize that there will be little or no evidence that the accuracy of the PCRA in predicting re-arrest depends on whether offenders are Black or White.
2. To what extent does the instrument yield average score differences between racial groups that are relevant to *disparate impact*? We hypothesize that Black offenders will obtain similar—or modestly higher—PCRA scores than Whites.
3. Which risk factors contribute the most and the least to mean score differences between Black and White offenders? We expect criminal history to contribute the most to these differences—and variable risk factors like substance abuse to contribute the least, in keeping with past research (Petersilia & Turner, 1987).
4. Are variables like criminal history best understood as proxies for race, or mediators of the relation between race and recidivism, given Kraemer et al.’s (2001) criteria? We hypothesize that the best classification will be “mediator.”

Our goal is to shed light on whether risk assessment has something to offer the justice system at this opportune moment for scaling back mass incarceration.

## METHOD

### Participants and Matching

Participants in this study were drawn from a population of 150,614 offenders who completed PCRA assessments as part of the probation intake process between August 2010 and November 2013 (see Walters & Lowenkamp, 2015). Offender eligibility criteria were: (a) assessed with the PCRA at least 12 months prior to the collection of follow-up arrest data (to permit tests of predictive bias;  $n$  lost = 83,894), (b) no missing data on PCRA items (to permit analyses at the risk factor level;  $n$  lost = 1,007), and (c) race coded as either “Black” or non-



Hispanic “White” (to permit relevant racial comparisons;  $n$  lost = 17,238). Application of these criteria yielded an eligible pool of 48,475 offenders. Given that even trivially small differences can become statistically significant in samples as large as ours (Lin, Lucas & Shmueli, 2013), we use an alpha level of .001 to signal statistical significance and focus on effect sizes in interpreting results. At this standard of  $p < .001$ , there were no significant differences between the eligible sample and the population from which it was drawn in age, sex, conviction offense, and PCRA total scores.

Within the eligible sample of 48,475 offenders, there were potentially confounding differences between Black and White participants. For example, Blacks were more likely to be young ( $d=0.44$ ) and male ( $d= .19$ ) than Whites (age and sex are robust risk factors for recidivism)—and the groups also differed in offense type (which can mark differential selection). To isolate the effect of race on risk and recidivism—without creating non-representative groups—we adopted a conservative matching approach.<sup>ii</sup> We randomly matched each Black offender to a White offender on age, sex, and offense using `ccmatch` in STATA (Cook, 2015). This process yielded a race-matched sample of 33,074 offenders. As shown in Table 1, the matched sample did not differ significantly at our standard of  $p < .001$  from the unmatched eligible sample across a range of characteristics. The prototypic offender was male, age 39, and convicted of a drug offense.

[Insert Table 1]

All offenders were followed for a minimum of one year, but the follow up period (i.e., time at risk for re-offending) was variable beyond that point. Compared to White offenders ( $M=1041$  days,  $SD=233$ ), Black offenders ( $M=1032$  days,  $SD=242$ ) had a significantly shorter follow-up period ( $t [33027.7] = -3.58; p < .001$ )—but the difference was just over one week, on

average ( $d=.04$ ). As shown later, our results include survival analyses that account for variable lengths of follow-up.

## **Measures of Risk**

The history, development, and predictive utility of the Post Conviction Risk Assessment (PCRA) are detailed elsewhere (see Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011; Lowenkamp et al., 2013; Lowenkamp, Holsinger, & Cohen, 2015). Briefly, the PCRA is an actuarial instrument that explicitly includes variable risk factors and was constructed and validated on large, independent samples of federal offenders. Items that most strongly predicted recidivism in the construction sample contribute most strongly to total scores. Fifteen items are scored and summed to yield a total PCRA risk score (Cronbach's  $\alpha=.71$ ) that places an offender into a risk category (low, low/moderate, moderate, or high). Each of the fifteen items is nested under one of five risk factor domains, four of which are changeable (i.e., all but criminal history). The domains and items are listed below. With the exception of the first two items listed, items are scored dichotomously (0 or 1):

- “Criminal history” includes number of prior arrests (0=none; 1=one-two; 2=three-six; 3=seven or more), young age (0=41+; 1=26-40; 2= under 26), community supervision violations, varied offending pattern, institutional adjustment problems, and violent offense ( $\alpha=.66$ ; Spearman-Brown Estimated  $\alpha$  |10 items=.76)
- “Employment and education” includes highest grade completed, unstable recent work history, and currently unemployed ( $\alpha=.47$ ; Spearman-Brown Estimated  $\alpha$  |10 items=.75)
- “Social networks” includes family problems, unmarried, and lack of social support ( $\alpha=.47$ ; Spearman-Brown Estimated  $\alpha$  |10 items=.67)

- “Substance abuse” includes recent alcohol problems and recent drug problems ( $\alpha=.38$   
Spearman-Brown Estimated  $\alpha$  |10 items=.80)
- “Attitudes” is low motivation to change

The PCRA has been shown to be reliable and valid. Specifically, officers must complete a training and certification process to administer the PCRA. The certification process has been shown to yield high rates of inter-rater agreement in scoring (Lowenkamp et al., 2012). The accuracy of the PCRA in predicting recidivism rivals that of other well-validated instruments (for a review, see Monahan & Skeem, 2014). For example, based on a sample of over 100,000 offenders, Lowenkamp et al. (2015) found that the PCRA moderately-to-strongly predicted both re-arrest for any crime and re-arrest for a violent crime, over up to a two-year period (AUCs=.70-.77). Finally, scores on the PCRA have been shown to change over time. Of offenders initially classified as high risk on the PCRA, 47% move to a lower risk classification upon reassessment an average of nine months later (Cohen & VanBenschoten, 2014). The greatest changes observed were in employment/education and substance abuse.

The PCRA was administered by agents when an offender entered supervision (within 90 days of intake), and takes 15-30 minutes to complete. In the present study, the results of the intake assessment were selected for analyses as this provided the longest follow up time period. In addition to the total PCRA score, the sub-scores from the PCRA domains (criminal history, education & employment, drugs & alcohol, social networks, and cognitions) were also calculated and used in some analyses.

### **Arrest Criterion**

Data from the National Crime Information Center (NCIC) and Access to Law Enforcement System were used to collect information on arrests. A standard criminal history

check was retrieved on each participant that yielded their entire criminal history. The date and types of arrests that occurred after the date of PCRA administration were coded from these data. The result was two dichotomous measures that we used in analyses of predictive fairness: arrest for any offense (excluding technical violations of standard conditions of supervision), and arrest for any violent offense. Violence was defined using the NCIC definitions (i.e., homicide and related offenses, kidnapping, rape and sexual assault, robbery, assault).

Our analyses and interpretation primarily focus on “violent arrest” because it is the most unbiased criterion available and “[c]onfidence in the criterion measure is a prerequisite for an analysis of predictive bias” (SIOP, 2003). According to differential selection theory, racial disparities reflect bias in policing and decisions about arrest. This theory applies less to crimes of violence than (victimless) crimes that involve greater police discretion (e.g., drug use, “public order” crimes; see Piquero & Brame, 2008). For the sake of completeness, we also report results for “any arrest.”

In our view, official records of arrest—particularly for violent offenses—are a valid criterion. First, surveys of victimization yield “essentially the same racial differentials as do official statistics. For example, about 60 percent of robbery victims describe their assailants as black, and about 60 percent of victimization data also consistently show that they fit the official arrest data” (Walsh, 2009, p. 22). Second, self-reported offending data reveal similar race differentials, particularly for serious and violent crimes (see Piquero, 2015). Third, changes in variable risk factors on the PCRA change the likelihood of future re-arrest (Cohen, Lowenkamp & VanBenschoten, 2015), suggesting that arrest statistics track risk-relevant behavior.

In the present sample, the base rate for any arrest was 27% (31% Black; 24% White,  $\chi^2(1) = 174.02$ ;  $p < 0.001$ ;  $\phi = -0.07$ ), and the base rate for violent arrest was 7% (9% Black; 6% White,

$\chi^2(1) = 94.46$ ;  $p < 0.001$ ,  $\phi = -0.05$ ). Although these base rates are not interpretable in an absolute sense because of the variable follow-up period, they indicate that Black participants were more likely to be arrested than White participants.

## **Analyses**

We calculated descriptive statistics, effect sizes, and measures of predictive validity. To test the PCRA's predictive fairness, we followed the standard practice of comparing the relative fit of specific nested regression models. Analyses are meant to represent the predictive fairness of PCRA scores in the federal population as a whole, across its 94 districts. To address concerns that the data may cluster by district, we used robust standard errors in the regression models to adjust for any heteroscedasticity. Specifically, the variance-covariance estimator with clustering by district was used to address the potential correlation between error terms within districts (STATA `vce[cluster]`; Guiterrez & Drukker, 2007; Rogers, 1993).

## **RESULTS**

### **Testing Predictive Fairness**

The first aim is to test the extent to which the PCRA—and the risk factors it includes—are free of predictive bias. We hypothesized that there will be little evidence that the accuracy of the PCRA in predicting re-arrest depends on whether offenders are Black or White. As shown below, results are generally consistent with this hypothesis.

**Strength of prediction.** First, we examined whether the *strength* or degree of relationship between PCRA total scores and re-arrest varied as a function of race. Table 2 presents re-arrest rates for offenders placed in each PCRA risk classification by race. Arrest rates increase monotonically as risk classifications increase, across racial groups.

[Insert Table 2]

Table 2 also presents DIF-R and AUC values by race. The Dispersion Index for Risk (DIFR; see Silver, Smith & Banks 2000) assesses the extent to which PCRA risk classifications create reasonably sized groups of offenders with maximally different arrest rates. DIFR ranges from 0 to infinity, increasing as the classification model disperses cases into groups whose base rates of arrest are distant from the total sample base rate and whose subgroup sizes are large in proportion to the total sample size. Unlike the DIFR (which focuses on PCRA risk classifications), the Area Under the ROC Curve (AUC) focuses on PCRA Total Scores. The AUC is an excellent measure of comparative predictive accuracy because its values are not influenced by base rates of offending (which vary across groups). Minimum AUCs of .56, .64, and .71 correspond to “small,” “medium,” and “large” effect sizes, respectively (see Rice & Harris, 1995).

As shown in Table 2, AUC values are consistently large, across racial groups. These values indicate, for example, a 72% (Black) or 75% chance (White) that an offender randomly selected from those who violently recidivated will obtain a higher PCRA score than an offender randomly selected from those who did not violently recidivate. The small AUC group differences reached statistical significance for any arrest ( $Z = -4.49$ ;  $p < 0.001$ ), but not violent arrest ( $Z = -2.47$ , *ns*). Similarly, DIFR values are consistently high across racial groups (see Skeem et al., 2013 for comparison), although values appear slightly higher for White participants.<sup>iii</sup>

**Form of prediction.** Having found that PCRA scores strongly predict arrest among both Black and White offenders, we next examined whether the *form* of the relationship between PCRA scores and recidivism varies as a function of race (Arnold, 1982). The crucial issue is whether an average PCRA score of  $X$  corresponds to an average arrest rate of  $Y$ , regardless of an

offender's race. The form of prediction (unlike its strength) is about the shape of the relationship between PCRA scores and recidivism by race.

To address this issue, we estimated a series of bivariate logistic regression models (four models for any arrest; four models for violent arrest). These models were compared to test for “subgroup differences in regression slopes or intercepts, [which] signal predictive bias” (SIOP, 2003). As shown in Table 3, in Models One and Two, only race and only the PCRA total score, respectively, were used to predict any arrest. Model Three included both race and the PCRA, and Model Four included race, the PCRA, and an interaction between race and PCRA. Each model was run using robust standard errors with clustering by district.

[Insert Table 3]

Model comparisons yielded two main findings. First, the slope of the relationship between PCRA scores and arrest is similar for Black and White offenders. That is, comparison of Models Three and Four indicate that the addition of the interaction term does not improve the prediction of any arrest [ $\chi^2(1) = 10.64, ns; Pseudo-R^2 \Delta=0.00$ ] or violent arrest, [ $\chi^2(1) = 0.28, ns; Pseudo-R^2 \Delta=0.00$ ]. The odds ratio for the interaction terms are also trivial and not statistically significant (see Table 3). In short, race does not moderate the utility of the PCRA in predicting any arrest or violent arrest. Second, there are no significant racial differences in the intercept of the relationship between PCRA total scores and any arrest, but the intercept of the relationship between PCRA scores and violent arrest is significantly lower for White than Black offenders. Specifically, comparison of Models Two and Three indicate that race adds no incremental utility to the PCRA in predicting any arrest [ $\chi^2(1) = 9.1, ns; Pseudo-R^2 \Delta=0.00$ ], but adds modest incremental utility in predicting violent arrest, [ $\chi^2(1) = 16.93, p <.001; Pseudo-R^2 \Delta=0.00$ ]. The odds ratios for race in Model Three are small and not statistically significant at our

standard of  $p < .001$ . Still, after taking PCRA scores into account, White offenders are 13% less likely to have a violent arrest than Black offenders ( $RR=0.83$ ). So there is modest overestimation of violent recidivism for White offenders.

In samples as large as ours, “almost any difference between models is likely to be statistically significant even if the difference has no practical importance” (Tabachnik & Fidell, 2007, p. 458). To concretize any racial differences in the form of the relation between the PCRA and any arrest, we (a) estimated the predicted probabilities of any re-arrest based on regression Model 4, (b) grouped those probabilities together for each PCRA score,<sup>iv</sup> and (c) displayed those grouped probabilities by race in Figure 1. Given the results above, one would expect—and one observes—that the two lines would be nearly identical. Across PCRA scores, predicted probabilities of arrest for Black and White offenders are highly similar in elevation and shape.

[Insert Figure 1]

**Supplemental analyses.** We tested the robustness of our results across four different dimensions. For the first three dimensions, we chiefly are interested in robustness for the most unbiased criterion available—“violent arrest.” The fourth and final dimension shifts focus to the potentially most biased criterion available—“any arrest or revocation.”

First, we wished to ensure that results were not confounded by variability in participants’ length of follow-up. To account for varying time at risk, while assessing whether race moderated the relationship between PCRA scores and recidivism, we completed sequential Cox regression analyses in which we entered race and PCRA scores in the first block, and then an interaction between race and PCRA scores in the second block, as predictors of either time to any arrest or violent arrest. After entering the first block, the addition of the second block reached statistical significance for any arrest [ $\Delta\chi^2(1) = 17.15, p < .001$ ], but not violent arrest [ $\Delta\chi^2(1) = 0.68, ns$ ].



The effect size for the interaction term of interest was small for both any arrest (OR=1.03,  $p < .001$ , 99.9% CI [1.01, 1.05]) and violent arrest (OR=1.01, *ns*, 99.9% CI [0.98, 1.06]). Compared to our regression-based results, these survival-based results are the same for violent arrest and similar for any arrest. This consistency suggests that our results are not confounded by varying lengths of follow-up. Flores et al.'s (in press) finding that variable- and fixed- follow up periods yield similar predictive estimates for the PCRA lend additional confidence to our findings.

Second, to ensure that our results were not a function of our approach to handling nested data (i.e., using robust standard errors with clustering), we completed a non-linear hierarchical model of Model 4, using HLM 7.01 analyses that clustered offenders within jurisdictions. The results were highly consistent with our main analyses. Specifically, PCRA Total scores significantly predicted violent arrest [OR=1.29,  $p < .001$ , 99.9% CI (1.25, 1.32)] and any arrest [OR=1.29,  $p < .001$ , 99.9% CI (1.27, 1.32)], but the remaining terms in the model did not [Race OR= 0.80, 99.9% CI (0.58, 1.22] & OR=.80 , 99.9% CI (0.62, 01.03]; Race x PCRA OR= 1.00, 99.9% CI (0.96, 1.04] & OR=1.02, , 99.9% CI (0.99, 1.05], for violent arrest & any arrest, respectively; all terms *ns*).

Third, to examine test fairness for factors that include both race and its risk-relevant correlates (e.g., age, gender, offense type), we completed the four core regression models with the eligible *unmatched* sample (N=48,475) for both violent arrest and any arrest. We obtained a similar pattern of results as with the matched sample. Specifically, comparison of Models Three and Four indicate that the addition of the interaction term significantly improved the prediction of any arrest [ $\chi^2(1) = 29.42$ ,  $p < .001$ ], but not violent arrest [ $\chi^2(1) = 4.54$ , *ns*, OR for interaction=1.03, *ns*, 99.9% CI (0.99, 1.07)]. For any arrest, the increase in explanatory power was trivial (*Pseudo-R*<sup>2</sup>  $\Delta=0.00$ ) and the interaction term was small (OR =1.04,  $p < .001$ , 99.9%

CI [1.01, 1.07]). Still, the PCRA's accuracy in predicting any arrest—but not the less biased criterion of violent arrest—may depend on race plus its risk-relevant correlates like age. The intercept of the relationship between PCRA scores and both violent arrest and any arrest was significantly lower for unmatched White than Black offenders [Model 2 vs. 3  $\chi^2(1) = 65.87$  &  $83.22, p < .001$ ; OR for race = 0.74, 99.9% CI (0.62, 0.87] & 0.81, 99.9% CI (0.71, 0.93)],  $p < .001$  for violent arrest and any arrest, respectively]; suggesting overestimation of arrest for White offenders.

Together, these results lend confidence to our main findings by indicating that they are not just a function of variable follow-up periods, nesting by jurisdiction, or sample matching to isolate the effects of race. Results for the most unbiased criterion available—violent arrest—were the same, for main- and supplemental- analyses. Next, we present a final series of analyses that test the robustness of our findings to potential criterion contamination.

Specifically, our fourth set of analyses explored whether test fairness generalizes from violent arrest to “any arrest or revocation.” This criterion is more subject to differential selection, given that it includes any arrest (see above, method) and probation revocations, which can be influenced by probation agents who are aware of offenders' PCRA scores and exercise discretion in their surveillance and reporting practices. Nevertheless, a reviewer observed that revocation may sometimes capture new offenses that are processed as revocations rather than arrests (as an easier way to get an offender “off the street”). So we completed the core set of four regression analyses using “any arrest or revocation” as the criterion—and obtained a similar pattern of results. Specifically, comparison of Models Three and Four indicate that the addition of the interaction term does not improve the prediction of any arrest or revocation [ $\chi^2(1) = 9.97, ns$ ; OR for interaction = 1.03,  $ns$ , 99.9% CI (0.99, 1.08)]. This indicates that the PCRA's accuracy in

predicting “any arrest or revocation” does not depend on race. There was also no significant difference between racial groups in the intercept of the relationship between PCRA scores and “any arrest or revocation” [Model 2 vs. 3  $\chi^2(1) = 3.304, ns$ ; OR for race= 0.97, *ns*, , 99.9% CI (0.84, 1.11)].

**Exploring predictive fairness at the risk factor level.** Even if there is little evidence of predictive bias at the global level for PCRA total scores, individual risk domains may be more- or less- racially fair in a manner that may be generalizable. To explore this possibility, we completed analyses that parallel those described above, to assess whether the relationship between each risk domain and any rearrest was similar in degree and form across race.

Table 4 shows the *degree* of association between PCRA domain scores and arrest, by race. As shown there, criminal history generally had a large effect in predicting arrest, and the remaining four domains had a small-medium effect. Criminal history, substance use, social networks predicted any arrest—but not violent arrest—better for White than Black participants. There were no other group differences.

[Insert Table 4]

Next, we assessed the predictive fairness of each PCRA risk factor. For each risk domain, we completed a series of four logistic regression models that parallel those described above for PCRA total scores (one series each for any arrest and violent arrest). Table 5 displays model comparisons that test for group differences in slopes and intercepts. Results indicate that race moderates the effect of substance use and social networks in predicting any arrest—but not violent arrest. In contrast, intercept differences were the rule rather than the exception: Criminal history was the only domain in which the intercept of the relationship between PCRA scores and

recidivism was similar for Black and White offenders. For other domains (especially substance use), PCRA scores tended to overestimate recidivism rates for White offenders.

[Insert Table 5]

**Summary.** Taken together, results are consistent with our hypothesis of predictive fairness by race. Specifically, the *form* of the relationship between PCRA total scores and re-arrest is very similar for Black and White offenders. There is a strong *degree* of relationship between PCRA total scores and re-arrest for both groups. Shifting from the global to the specific level, the substance abuse and social network domains predicted any arrest better for White than Black offenders; but there was little evidence of predictive bias *per se* for the remaining domains. Any domain-level differences tended to overestimate recidivism for White participants.

#### **Assessing Mean Score Differences Relevant to Disparate Impact**

**Matched sample.** The second aim was to assess the extent to which racial groups obtain different scores on the PCRA relevant to *disparate impact*. We hypothesized that Black offenders would obtain similar—or modestly higher—PCRA scores than Whites. The mean PCRA total score was 7.37 ( $SD= 3.25$ ) for Black participants and 6.23 ( $SD= 3.38$ ) for White participants—an average 1.1-point difference on an 18-point scale. The effect of race on PCRA scores is  $d= .34$ , which translates to 13.5% non-overlap—and 86.5% overlap—between racial groups in PCRA scores (see Reiser & Faraggi, 1999).

**Supplemental results for unmatched sample.** The results described above isolate the effect of race on PCRA scores, excluding the correlated effects of age, gender, and offense type. To supplement these results, we also calculated mean score differences for the eligible *unmatched* sample ( $N=48,475$ ). There was an average 1.9-point difference in PCRA total scores in this sample: Scores were 7.65 ( $SD=3.21$ ) for Black participants and 5.79 ( $SD= 3.45$ ) for

White participants. The effect of race on PCRA scores is  $d = .56$  (CI=.53-.58), which translates to 22% non-overlap—and 78% overlap—between Black and White groups in PCRA scores.

### **Identifying Risk Factors That Underpin Mean Score Differences**

**Domain differences.** Our third aim was to determine which risk factors contribute the most to mean score differences between Black and White offenders. We expected criminal history to contribute the most—and variable risk factors like substance abuse and attitudes to contribute the least. Results are consistent with this hypothesis.

Mean scores and standard deviations for PCRA risk domains (and total scores) are reported by race in the upper panel of Table 6, along with Cohen's  $d$ . We include the percentage of the difference in the PCRA total means that is attributable to a given risk domain. As shown in Table 6, 66% of the racial difference in mean PCRA scores is attributable to differences in criminal history (this figure rises to 73% in the unmatched sample). Most of the remaining difference (28%) is attributable to the employment and education domain. The effect of race on criminal history ( $d = .34$ ) and employment/education ( $d = .33$ ) is essentially the same as that of total PCRA scores. The remaining three PCRA domains—substance abuse, attitudes, and social networks—contributed negligibly to mean score differences between Black and White offenders.

[Insert Table5]

**Drilling down on criminal history.** Because criminal history can be measured in myriad ways, Frase et al. (2015) recommend that individual items be examined by race. In the lower panel of Table 5, we display mean score differences by race for five of the six criminal history items (age is omitted because the sample was age-matched). The effect of race for each criminal history item is similar, with the number of prior arrests ( $d = .41$ ) and past violent offenses ( $d = .36$ ) accounting for the majority of the difference in criminal history scores.

## Proxy or Mediator?

Finally, we assess whether criminal history is a proxy for race or a mediator of the relation between race and recidivism. We focus on violent arrest, the most unbiased criterion.

In determining the relationship between two risk factors (in this case, A=race and B=criminal history), Kraemer et al (2001) focus on three elements: temporal precedence (of A and B, which comes first?); correlation (are A and B correlated?); and dominance (would the use of A alone, B alone, or one of the two combinations of A and B—i.e., A and B; A or B—yield greatest potency in predicting arrest?). Applying these criteria, race precedes criminal history and race and criminal history are correlated ( $r = -.17$ ). Criminal history is not a proxy for race, however, because race does not “dominate” in predicting violent arrest: Instead, criminal history ( $r_p = .21$ ) predicts violent arrest more strongly than race ( $\phi = -.05$ ).

Following Kraemer et al.’s framework, then, criminal history mediates the relationship between race and future violent arrest. To assess whether criminal history fully mediates or partially mediates this relationship (i.e., whether criminal history dominates race, or criminal history and race co-dominate), we completed a series of mediation analyses using the `binary_mediation` package in STATA (Ender, 2011). This package combines linear regression with logit models to calculate indirect effects of mediator variables (binary or continuous) on a response variable (binary or continuous), using standardized coefficients and a product of coefficients approach. Standard errors and confidence intervals are generated through bootstrapping. Results are consistent with partial mediation. Specifically, after controlling for criminal history, race was a weak, but still statistically significant predictor of violent arrest  $b = -.09, p < .001$ . Both the direct coefficient ( $b = -.09, SE = .03, p < .001$ ), and the indirect coefficient

were significant ( $b = -.29$ ,  $SE = .01$ ,  $p < .001$ ). However, 76% of the total effect of race on future violent arrest was mediated by criminal history.

### **Putting Predictive Fairness and Mean Score Differences Together**

In Figure 2, we provide a visual summary of the study's global findings. In this figure, PCRA scores appear on the X axis. The number of offenders (0-2,000) appear on the right Y axis and arrest rates (0-100%) appear on the left Y axis. The figure shows (a) the area of non-overlap between Black and White groups in PCRA distributions (much of it falling at the low end), and (b) the similar increase in arrest rates for Black and White offenders across the PCRA scale.

## **DISCUSSION**

At the most basic level, these results indicate that risk assessment is not “race assessment.” First, there is little evidence of test bias for the PCRA. The instrument strongly predicts re-arrest for both Black and White offenders. Regardless of group membership, a PCRA score has essentially the same meaning, i.e., same probability of recidivism. So the PCRA is informative, with respect to utilitarian and crime control goals of sentencing. Second, Black offenders tend to obtain higher scores on the PCRA than White offenders ( $d = .34$ ; 13.5% non-overlap). So some applications of the PCRA might create disparate impact—which is defined by moral rather than empirical criteria. Third, most (66%) of the racial difference in PCRA scores is attributable to criminal history—which strongly predicts recidivism for both groups, is embedded in current sentencing guidelines, and has been shown to contribute to disparities in incarceration (Frase et al., 2015). Finally, criminal history is *not* a proxy for race. Instead, criminal history partially mediates the weak relationship between race and a future violent arrest.

Are these results merely a function of “bias predicting bias,” e.g., biased criminal history records predicting biased future police decisions about arrest? Put more broadly, is the

appearance of validity for the PCRA due to differential selection? In a word—no. First, criminal history predicts violent arrest with similar strength and form, whether participants are Black or White (Table 4). Second, the PCRA’s power in predicting arrest is not explained by criminal history. That is, after controlling for criminal history scores ( $OR = 1.48, p < .001, 99.9\% CI [1.41, 1.56]$ ), PCRA “need” scores (i.e., employment-education, social networks, substance abuse, and attitudes;  $OR = 1.18, p < .001, 99.9\% CI [1.14, 1.22]$ ) add significant incremental utility in predicting arrests for violence for both Black and White participants,  $\Delta\chi^2(1) = 132.57, p < .001$ . Third, risk assessment instruments like the PCRA have been shown to predict not only official records of arrest, but also self-reported and collateral-reported offending (Monahan et al., 2001; Yang et al., 2010). Together, these facts (and others) rule out the possibility that these findings are mere artifacts of differential selection.

Before unpacking our findings, we note four study limitations that must be borne in mind. First, we used a sample of Black and White offenders matched in age, gender, and offense type. Because this study is among the first to focus on the topic, we wished to isolate the effects of race. As shown above, parallel analyses completed with the eligible (non-matched) sample yielded the same results for violent arrest. Second, our results may not generalize beyond the federal system. The PCRA was specifically developed for federal offenders, who differ from state-level offenders. For example, although the PCRA strongly predicts future violent arrests (Table 2), federal offenders are much less likely to have been convicted of violent offenses than state offenders (Carson, 2015). Third, interrater reliability data on the PCRA are not available for the present sample, although all officers must complete a PCRA certification process that has been shown to yield reliable scores (Lowenkamp et al., 2013). Fourth, as is the case in most studies of this kind, probation services and supervision may have affected participants’



recidivism rates. To confound our main findings, however, services would have to be more effective for Black than White participants, which seems unlikely (e.g., Lipsey et al., 2007 found that race did not significantly moderate the effect of evidence-based treatment on recidivism).

### **Little Evidence of Test Bias**

The degree and form of association between PCRA total scores and arrest were similar, for Black and White offenders. These findings are consistent with past studies indicating that the *degree* of association between other “risk-needs” tools and recidivism are similar for Black and White offenders (Brennan et al., 2009; Lowenkamp & Bechtel, 2007; Kim, 2010). But we went beyond past research to test whether the *form* of the relationship between risk and recidivism is similar across races. In Figure 1, we show that a given PCRA score has similar meaning, regardless of group membership. There were no meaningful differences between Black and White offenders in slopes of the relationships between PCRA scores and future arrests—and the one difference observed for the intercept of this relationship conveys modest overestimation for White offenders (e.g., of PCRA-classified moderate risk offenders, rates of violent arrest are 14% and 16% for White and Black offenders, respectively; Table 1).

The appropriate level for assessing test fairness is the test level—not the subscale level. However, having established little predictive bias for PCRA total scores, we also examined specific risk factors—some of which have been labeled as racially unfair by critics (i.e., criminal history and employment/education; Harcourt, 2015; Starr, 2014). For three of the five risk domains—including those claimed to be biased—there was no evidence that race moderated their predictive utility. Slope differences were evident for only two factors—i.e., recent substance abuse problems and social networks—which predicted any arrest, but not violent arrest, more strongly for White than Black offenders. This may indicate that the PCRA’s

definition of these risk constructs do not completely overlap across groups. For example, one of the PCRA's three "social network" domain items— "unmarried"—may be more common and therefore less indicative of social network problems for Black than White offenders (see Bureau of Labor Statistics, 2013; van de Vijver & Tanzer, 2004). The fact that some subscale-level bias did not translate to PCRA-level bias is consistent with the cognitive testing literature, where it is "common to find roughly equal numbers of differentially functioning items favoring each subgroup, resulting in no systematic bias at the test level" (SIOP, 2003, p. 34).

In summary, PCRA scores are useful for assessing risk of future crime, whether an offender is Black or White. The generalizability of these results to other risk assessment instruments is unclear. Risk assessment instruments that are very short, narrow in content, and/or developed with homogeneous samples may be more prone to bias than the PCRA.

### **Mean Score Differences Relevant to Disparate Impact**

**Size of race difference.** Mean score differences between groups are uniformly rejected as an indicator of test bias because group differences may reflect real differences. For example, the average weight of females is less than that of males, but this is not an indicator of scale bias. Still, mean score differences are relevant to disparate impact associated with the *use* of a test—and Black offenders are already incarcerated at a much greater rate than White offenders.

In the matched sample, the effect of race on PCRA scores was  $d = .34$ , which corresponds to 13.5% non-overlap—and 86.5% overlap—between Black and White groups. In the unmatched sample, the effect of race and its correlates (age, gender, and offense type) on PCRA scores was  $d = .56$ , which corresponds to 20% non-overlap and 80% overlap between groups. Cohen (1988) reluctantly provided benchmarks for interpreting  $d$  in behavioral research (i.e., .20=small/not

trivial; .50=medium; .80=large)—but strongly cautioned that “this is an operation fraught with many dangers” (p. 22). Effect sizes must be interpreted in light of past relevant findings.

On that note, the effect of race on PCRA scores is similar to the effect of race on criminal history scores embedded in sentencing guidelines ( $d = .19-.29$ ; or 8-12% non-overlap; data from Frase et al., 2015). More broadly, the effect of race on PCRA scores is smaller than that observed for high stakes cognitive tests. The results of a meta-analysis indicate a sizable effect of race on the SAT ( $d = 0.99$ ), ACT ( $d = 1.02$ ) and GRE ( $d = 1.34$ ; Roth, Bevier, Bobko, Switzer & Tyler, 2001). These effect sizes correspond to 38-51% non-overlap between Black and White groups.

There are no set criteria for determining when mean score differences are large enough to translate into disparate impact. First, inequitable social consequences—or “lack of fairness—is a social rather than psychometric concept. Its definition depends on what one considers to be fair” (SIOP, 2003, p. 31). Second, disparate impact is determined by the *use* of the instrument (not the instrument itself). Inequitable consequences may depend less on the magnitude of group differences in scores than on how those scores are used—i.e., what decision they inform, how heavily they are weighed, and what practices they replace.

Even uses of instruments that seem disconnected from racial disparities in incarceration can invoke definitions of fairness. For example, the PCRA is used strictly to inform risk reduction efforts, so one could argue that disparate impact is not an issue—if anything, Black people might be privileged for costly services designed to improve re-entry success. But those with a different view of fairness could argue that risk reduction efforts are not about service access, but about social control—more surveillance and more conditions of supervised release (see Swanson et al., 2009). When federal probationers are found to violate conditions (including treatment conditions), judges may “revoke a term of supervised release, and require the

defendant to serve in prison all or part of the term of supervised release...without credit for time previously served on postrelease supervision” (17 USC §3583(e)3). Of course, this view must be juxtaposed against a long tradition of relying upon risk assessment as a factor in probation, parole, and other accelerated release practices designed to use correctional resources efficiently.

In an effort to begin addressing nebulous issues around disparate impact, some states have adopted “Racial Impact Statement policies,” which “require an assessment of the projected racial and ethnic impact of new policies prior to adoption. Such policies enable legislators to assess any unwarranted racial disparities that may result from new initiatives and to then consider whether alternative measures would accomplish the relevant public safety goals without exacerbating disparities” (The Sentencing Project, 2000, p. 58).

**Differences chiefly attributable to criminal history.** Although disparate impact defies empirical definition, it is easy to objectively identify risk factors that contribute more- and less- to mean score differences between groups. Criminal history accounts for two-thirds of the racial difference in PCRA scores—partly because of its effect size and partly because this scale is weighed most heavily in total scores (i.e., contributes 9 of 18 possible points). As Frase et al. (2015) observe, the magnitude of racial differences in criminal history scores varies as a function of how sentencing guidelines operationalize this variable.

Criminal history presents a conundrum (Petersilia & Turner, 1987). On one hand, criminal history is among the strongest predictors of arrest and is perceived as relevant to an offender’s blameworthiness for the conviction offense (Monahan & Skeem, in press)—which may explain why criminal history has quietly become embedded in many jurisdictions’ sentencing guidelines, unlike other risk factors perceived as irrelevant to blameworthiness. On the other hand, heavy

reliance on criminal history at sentencing will contribute more to disparities in incarceration than reliance upon other robust risk factors less bound to race.

Although these concerns about criminal history are loosely consistent with Harcourt's (2015) criticisms, criminal history is not a proxy for race (as Harcourt contends). It is not the case that the principal connection between criminal history and arrest is race. Criminal history is better construed as a mediator, by Kraemer et al.'s (2001) criteria. We cannot infer causality from associations, but our results are consistent with what we would expect to see if a causal path leading from race to criminal history to violent future arrest were in force.

Our results are less consistent with Starr's (2014) objections to risk assessment. The employment/education domain was equally predictive of recidivism for Black and White offenders and accounted for only one-third of the racial difference in PCRA total scores. Moreover, employment/education—as operationalized in the PCRA—has been found to change over relatively short periods of time: Among high-risk offenders, 79% were unemployed and 87% lacked a stable recent work history at their initial assessment, compared to 49% and 66%, respectively, at their second assessment (Cohen & VanBenschoten, 2014). Although unrelated to blameworthiness, this risk factor is partly within an individual's control.

Differences between Black and White offenders across the remaining PCRA risk domains—social networks, substance abuse, and attitudes—were limited ( $d = -.04-.11$ ). This is broadly consistent with the view that variable risk factors are less objectionable than “static” and “immutable” characteristics. However, whether most variable risk factors are *causal*—i.e., would reduce recidivism if deliberately changed—is an open question that must be answered to inform risk reduction efforts (see Monahan & Skeem, in press).

**Familiar dilemma.** As an instrument, the PCRA is essentially free of predictive bias, but there are mean score differences between Black and White offenders that could translate into disparate impact. This dilemma is familiar in the cognitive testing domain, where mean score differences between Black and White groups are much larger than those observed here:

“Particularly with regard to race and ethnicity, the differences are of a magnitude that can result in substantial differences in selection or admission rates if the test is used as the basis for decisions. Employers and educational institutions wanting to benefit from the predictive validity of these tests but also interested in the diversity of a workforce or an entering class encounter the tension between these validity and diversity objectives. A wide array of approaches has been investigated as potential mechanisms for addressing this validity–diversity trade-off” (Sackett et al., 2008, p. 222).

Here, the issue is that risk assessment instruments can scaffold efforts to unwind mass incarceration without compromising public safety. But some applications of instruments might exacerbate racial disparities in incarceration. If one concern—predictive accuracy or social justice—is valued to the exclusion of the other, there is no dilemma. But if both concerns are valued—which is most likely—the two goals must be balanced (see Sackett et al., 2001).

### **Implications**

This study’s most straightforward implication is that risk assessment instruments should be routinely tested for predictive bias and mean score differences by race. For obvious reasons, these are fundamental standards of testing—particularly in high stakes domains (see The Standards, Section 3). We recommend that these issues be examined not only at the test level, but also at the level of risk factors. If policymakers blindly eradicate risk factors from a tool because they are contentious, they risk reducing predictive utility *and* exacerbating the racial

disparities they seek to ameliorate. It may be politically tempting, for example, to focus an instrument tightly on criminal history because this variable is associated with perceptions of blameworthiness, and is also easily assessed by referring to conviction records. But risk estimates based on a broader set of factors predict recidivism better than criminal history and tend to be less correlated with race (e.g., Berk 2009).

As suggested above, a number of strategies have been tested for maximizing an instrument's predictive utility while minimizing mean score differences. For example, in the context of selection for employment and education, efforts have been made to identify other predictors of work- and academic- performance (e.g., personality, interests, socioemotional skills; Sackett et al., 2001). Reasoning by analogy, efforts could be undertaken in the risk assessment domain to rely less heavily on criminal history while weighting risk factors with fewer mean score differences more heavily. Whether and how such strategies will “work” is unclear—but this is an important empirical question that we are now addressing.<sup>v</sup>

## **Conclusion**

In light of our results, it seems that concerns expressed about risk assessment are exaggerated. To be clear, we are not offering a blanket endorsement of the use of risk assessment instruments to inform sentencing. There will always be bad instruments (e.g., tests that are poorly validated) and good instruments “used inappropriately (e.g., tests with strong validity evidence for one type of usage put to a different use for which there is no supporting evidence)” (Sackett et al., 2008, p. 225). We are simply offering a framework for examining important concerns related to race, risk assessment, and recidivism. Our results demonstrate that risk assessment instruments *can* be free of predictive bias and *can* be associated with small mean

score differences by race. They also provide some direction for improving instruments in a manner that might balance concerns about predictive utility and disparate impact.

This article focuses on one factor that would influence whether the use of risk assessment in sentencing would exacerbate, mitigate, or have no effect on racial disparities in imprisonment—the instrument itself. But the instrument is only part of the equation. Given findings in the general sentencing literature, the effect of risk assessment on disparities will also vary as a function of the baseline sentencing context: Risk assessment, compared to what? Racial disparities depend on where one is sentenced (Ullmer 2012), so—holding all else constant—the effect of a given instrument on disparities will depend on what practices are being replaced (Monahan & Skeem, in press; see also Ryan & Ployhart, 2014).

Although practices vary, common denominators include (a) judges' intuitive consideration of offenders' likelihood of recidivism, which is less transparent, consistent, and accurate than evidence-based risk assessment (see Rhodes et al., 2015), and (b) sentencing guidelines that heavily weight criminal history and have been shown to contribute to racial disparities (Frase 2009). There is at least one demonstration that risk assessment does not lead to more punitive sentences for high-risk offenders (albeit in the Netherlands; see van Wingerden, van Wilsem, & Moerings, 2014). There is no empirical basis for assuming that the status quo—across contexts—is preferable to judicious application of a well-validated and unbiased risk assessment instrument. We hope the field proceeds with due caution.



Table 1: Sample Characteristics

Characteristic	Eligible Unmatched Sample (N=48,475)	Race-Matched Sample (N=33,074)
PCRA Total Score	6.74	6.81
Age	39.99	39.39
% White	48.62	50.00
% Male	85	84
% Conviction offense <sup>^</sup>		
Drug	46	47
Firearms	16	16
White Collar	17	18
Other	8	9
Violence	5	5
Property	5	5

<sup>^</sup> Categories with less than 5% combined as other (i.e., sex offense, public order)

PCRA=Post Conviction Risk Assessment

Table 2. Predictive Utility of PCRA by Race

Feature	Any Arrest			Violent Arrest		
	All	Black	White	All	Black	White
% Arrested by PCRA Classification						
Low	11	12	10	2	2	2
Low/Moderate	29	30	27	7	8	7
Moderate	49	49	48	15	16	14
High	64	62	66	21	23	19
DIF-R, PCRA Categories	0.83	0.78	0.85	0.99	0.91	1.01
AUC, PCRA Total	0.73	0.71	0.74	0.74	0.72	0.75

Note: N=33,074. PCRA= Post Conviction Risk Assessment; DIF-R= Dispersion index; AUC=Area Under the ROC Curve

Table 3. Logistic Regression Models Testing Predictive Fairness of PCRA by Race

	Any Arrest							
	Model 1	99.9% CI	Model 2	99.9% CI	Model 3	99.9% CI	Model 4	99.9% CI
Race (White)	0.72*	0.66, 0.78	--	--	0.92	0.84, 1.01	0.73	0.52, 1.02
PCRA Total	--	--	1.30*	1.29, 1.32	1.30*	1.28, 1.32	1.28*	1.26, 1.31
Race * PCRA Total	--	--	--	--	--	--	1.03	1.00, 1.06
(Constant)	0.44*	0.42, 0.47	0.05*	0.05, 0.06	0.06*	0.05, 0.06	0.06*	0.05, 0.07
Model $\chi^2$	62.79*		2133.88*		2201.96*		2378.53*	
Model Pseudo- $R^2$	0.01		0.11		0.11		0.11	
	Violent Arrest							
	Model 1	99.9% CI	Model 2	99.9% CI	Model 3	99.9% CI	Model 4	99.9% CI
Race (White)	0.66*	0.57, 0.76	--	--	0.83	0.69, 1.01	0.78	0.48, 1.26
PCRA Total	--	--	1.29*	1.27, 1.32	1.29*	1.26, 1.32	1.29*	1.25, 1.33
Race * PCRA Total	--	--	--	--	--	--	1.01	0.96, 1.06
(Constant)	0.09*	0.09, 0.10	0.01*	0.01, 0.01	0.01*	0.01, 0.01	0.01*	0.01, 0.02
Model $\chi^2$	52.21		1602.32		1691.89		1676.94	
Model Pseudo- $R^2$	0.001		0.09		0.09		0.09	

\*  $p < .001$

Note: Values for predictors are odds ratios, with race terms representing the unique effect for White compared to Black (i.e., White dummy coded as 1). N=33,074

Table 4. Utility of PCRA Domain Scores in Predicting Arrest by Race

	Any Arrest, AUCs			Violent Arrest, AUCs		
	All	Black	White	All	Black	White
Criminal History	0.71*	0.69	0.73	0.73	0.71	0.75
Employment	0.62	0.61	0.62	0.62	0.62	0.61
Drugs/Alcohol	0.58*	0.56	0.60	0.57	0.57	0.58
Social Networks	0.60*	0.58	0.61	0.59	0.59	0.60
Attitude	0.55	0.55	0.55	0.55	0.55	0.54

Note: AUC=Area under the ROC curve

\* differences significant at  $p < .001$  for any arrest (no significant differences for violent arrest)

Table 5. Logistic Regression Models Testing Racial Fairness of PCRA Domains in Predicting Arrest

	Slope Comparisons (Models 3 vs. 4)				Intercept Comparisons (Models 2 vs. 3)			
	R <sup>2</sup> Change	X <sup>2</sup>	OR, Interaction (Model 4)	99.9% CI	R <sup>2</sup> Change	X <sup>2</sup>	OR, Race (Model 3)	99.9% CI
<b>Any Arrest</b>								
Criminal History	0.00	5.27	1.03	0.97, 1.10	0.00	12.85*	0.91	0.79, 1.05
Employment	0.00	4.21	1.05	0.96, 1.16	0.00	56.44*	0.83*	0.72, 0.94
Drugs/Alcohol	0.00	31.53*	1.29*	1.10, 1.51	0.01	205.31*	0.69*	0.61, 0.80
Social Networks	0.00	17.94*	1.01*	1.02, 1.28	0.00	145.45*	0.74*	0.65, 0.84
Attitudes	0.00	5.25	1.12	0.94, 1.47	0.00	142.39*	0.74*	0.65, 0.85
<b>Violent Arrest</b>								
Criminal History	0.00	1.85	1.03	0.94, 1.14	0.00	14.67*	0.84	0.70, 1.02
Employment	0.00	0.017	0.99	0.86, 1.15	0.00	39.85*	0.76*	0.62, 0.92
Drugs/Alcohol	0.00	0.73	1.05	0.82, 1.33	0.01	105.63*	0.64*	0.53, 0.77
Social Networks	0.00	1.23	1.06	0.89, 1.25	0.00	82.44*	0.67*	0.56, 0.82
Attitudes	0.00	0.44	1.08	0.49, 1.47	0.00	81.40*	0.68*	0.56, 0.82

Note: OR=Odds Ratio, with terms representing the unique effect for White compared to Black (White dummy coded 1); N=33,074

\* $p < .001$

Table 6. PCRA Mean Score Differences by Race

Variable	Black (N=16,537)		White (N=16,537)		Difference	% Attributable To	Cohen's d		
	Mean	Std. Dev.	Mean	Std. Dev.			Estimate	Lower	Upper
PCRA Total	7.37	3.25	6.23	3.38	1.14		0.34	0.31	0.36
<u>Domains</u>									
Criminal History	4.74	2.16	4.00	2.28	0.75	66	0.34	0.32	0.37
Employment/Education	1.15	1.01	0.84	0.92	0.32	28	0.33	0.31	0.35
Substance Abuse	0.22	0.50	0.25	0.53	-0.03	-3	-0.06	-0.08	-0.04
Social Networks	1.12	0.79	1.05	0.79	0.07	6	0.09	0.07	0.11
Attitudes	0.13	0.34	0.10	0.29	0.04	3	0.11	0.09	0.13
Criminal History Domain	4.74	2.16	4.00	2.28	0.75		0.34	0.32	0.37
<u>Items</u>									
Prior Arrests	2.01	1.02	1.69	1.09	0.32	43	0.30	0.28	0.32
Violent Offenses	0.53	0.50	0.38	0.49	0.15	20	0.31	0.28	0.33
Varied Offending	0.77	0.42	0.67	0.47	0.10	13	0.22	0.20	0.24
Conditional Sup'n Violation	0.49	0.50	0.39	0.49	0.09	13	0.19	0.17	0.21
Institutional Adjustment	0.26	0.44	0.19	0.39	0.08	10	0.19	0.17	0.21

Note: PCRA= Post Conviction Risk Assessment

Figure 1. Predicted Probabilities of Arrest by PCRA Score and Race

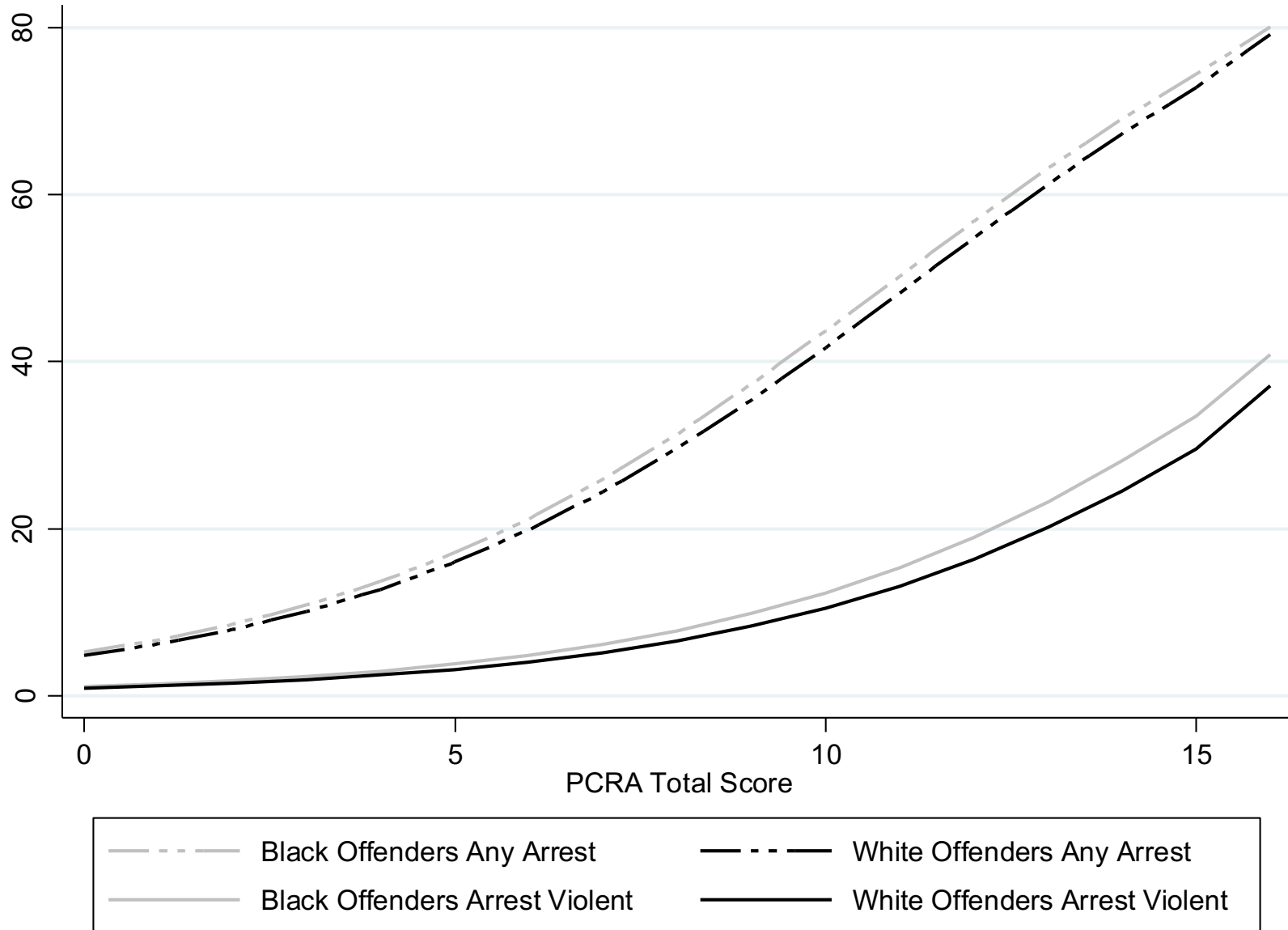
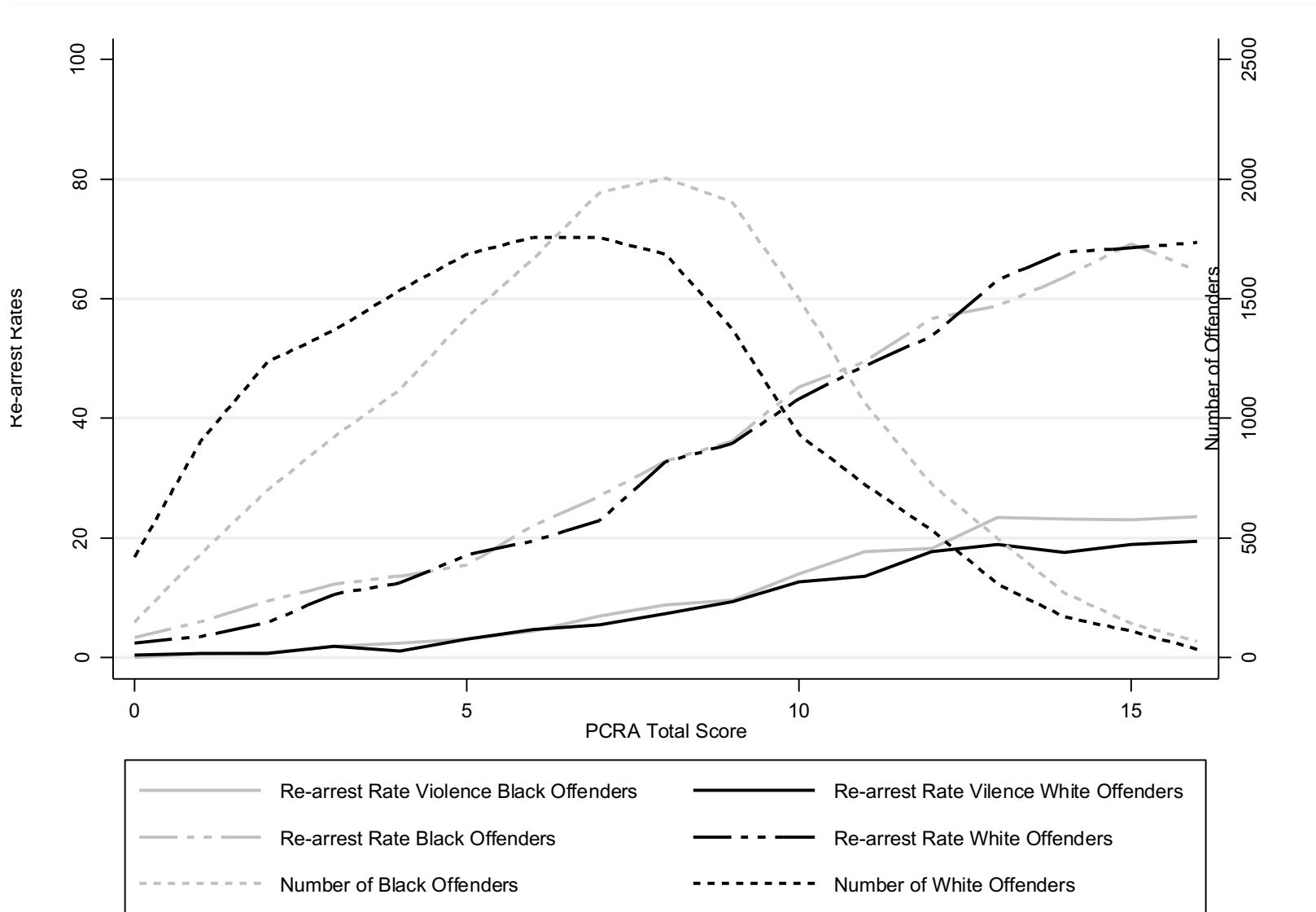


Figure 2. Rate of Arrest and PCRA Distribution by Race





## REFERENCES

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *The Standards for Educational and Psychological Testing*. Washington, DC: AERA Publications.
- American Law Institute (2014). *Model Penal Code: Sentencing (Tentative Draft No. 3)*. Philadelphia: American Law Institute.
- Arnold, H. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior & Human Performance*, 29 143-174.
- Arnold J, Arnold L. 2015. Fixing justice in America. *Politico Magazine*.  
<http://www.politico.com/magazine/story/2015/03/criminal-justice-reform-coalition-for-public-safety-116057.html>
- Berk, R. (2009). The role of race in forecasts of violent crime. *Race and social problems*, 1, 231-242.
- Blumstein, A. (1993). Racial disproportionality of US prison populations revisited. *University of Colorado Law Review*, 64, 743-760.
- Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36, 21-40.
- Bureau of Labor Statistics (October, 2013). Marriage and divorce: Patterns by gender, race, and educational attainment. Retrieved 10/10/15 from:  
<http://www.bls.gov/opub/mlr/2013/article/marriage-and-divorce-patterns-by-gender-race-and-educational-attainment.htm>
- Carson, E. A. (2015). Prisoners in 2014. Washington, DC: Bureau of Justice Statistics. Retrieved 10/10/15 from: <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5387>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2<sup>nd</sup> ed. New Jersey: Lawrence Erlbaum.
- Cohen, T, Lowenkamp, C, & VanBenschoten, S (2015). Does Change in Risk Matter? Examining Whether Changes in Offender Risk Characteristics Influence Recidivism Outcomes. Available at SSRN: <http://ssrn.com/abstract=2621267>
- Cohen, T. H., & VanBenschoten, S. W. (2014). Does the risk of recidivism for supervised offenders improve over time? Examining changes in the dynamic risk characteristics for offenders under federal supervision. *Federal Probation*, 78, 41-52.
- Cook, D.E. (2015). CCMATCH: Stata module to randomly match cases and controls based on specified criteria. Version 1.3. [www.Danielecook.com](http://www.Danielecook.com) .

- Desmarais, S.L., Johnson, K.L., & Singh, J.P. (2015). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services*.
- Durose, M., Cooper, A., & Snyder, H. (2014). *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*. Washington, D.C.: Bureau of Justice Statistics.
- Elek, Warren, R. & Casey, (2015). Using Risk and Needs Assessment Information at Sentencing: Observations from Ten Jurisdictions. Williamsburg, VA: National Center for State Courts. Available: <http://www.ncsc.org/~media/Microsites/Files/CSI/RNA%202015/Final%20PEW%20Report%20updated%2010-5-15.ashx>
- Ender, P.B. (2011). Binary\_mediation: Command to compute indirect effect with binary mediator and/or binary response variable. UCLA: Statistical Consulting Group. Available: <http://www.ats.ucla.edu/stat/stata/ado/analysis/>.
- Flores, A., Holsinger, A., & Lowenkamp, C. (in press). Comparing variable and fixed follow-up outcome periods: Do different methods produce different results? *Criminal Justice & Behavior*.
- Frase, R. S. (2004). Limiting retributivism. In M. Tonry (Ed), *The Future of Imprisonment*. New York: Oxford University Press.
- Frase, R. S. (2009). What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations? *Crime and Justice*, 38, 201-280.
- Frase RS. (2013). *Just Sentencing: Principles and Procedures for a Workable System*. New York: Oxford Univ. Press
- Frase, R.S. (2014). Recurring policy issues of guidelines (and non-guidelines) sentencing: Risk assessments, criminal history enhancements, and the enforcement of release conditions. *Federal Sentencing Reporter*, 26, .145-157.
- Frase, R.S., Roberts, J.R., Hester, R. & Mitchell, K.L. (2015). *Criminal History Enhancements Sourcebook*. Minneapolis, MN: Robina Institute of Criminal Law and Criminal Justice. Available: <http://www.robinainstitute.org/publications/criminal-history-enhancements-sourcebook/>
- Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works!. *Criminology*, 34, 575-608.
- Gottfredson, M. R., & Gottfredson, D. M. (1988). *Decision Making in Criminal Justice: Toward the Rational Exercise of Discretion*, 2<sup>nd</sup> ed. New York: Plenum Press.

- Griggs v. Duke Power Co.* (1971) 401 U.S. 424
- Guitterez, R & Drukker, D. (2007). Stata's cluster-correlated robust variance estimates. Available: <http://www.stata.com/support/faqs/statistics/references/>
- Harcourt, B. E. (2008). *Against prediction: Profiling, policing, and punishing in an actuarial age*. Chicago, IL: University of Chicago Press.
- Harcourt, B. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter* 27: 237-243.
- Hoge, R. D. (2002). Standardized instruments for assessing risk and need in youthful offenders. *Criminal Justice and Behavior*, 29, 380-396.
- Holder, E. (2014). Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting. Available at: <http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th>
- Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, 75, 16-29.
- Kim, H. S. (2010). *Prisoner classification re-visited: A further test of the Level of Service Inventory-Revised (LSI-R) intake assessment* (Doctoral dissertation, Indiana University of Pennsylvania).
- Kleiman, M., Ostrom, B., & Cheesman, F. (2007). Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, 53, 106-132.
- Kraemer, H.C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry* 158:848–856
- Lawrence A. 2013. Trends in Sentencing and Corrections: State Legislation. Denver: National Conference of State Legislatures  
<http://www.ncsl.org/Documents/CJ/TrendsInSentencingAndCorrections.pdf>
- Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4), 906-917.

- Lowenkamp, C. T., & Bechtel, K. (2007). Predictive Validity of the LSI-R on a Sample of Offenders Drawn from the Records of the Iowa Department of Correction Data Management System. *Federal Probation, 71*, 25-34.
- Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA Revisited: Testing the Validity of the Federal Post Conviction Risk Assessment (PCRA). *Psychological Services, 12*, 149-157.
- Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The Federal Post Conviction Risk Assessment (PCRA): A construction and validation study. *Psychological Services, 10*, 87-96.
- McCord, J., Widom, C. S., & Crowell, N. A. (2001). *Juvenile Crime, Juvenile Justice. Panel on Juvenile Crime: Prevention, Treatment, and Control*. Washington, DC: National Academy Press.
- Monahan, J., & Skeem, J. (2014). Risk redux: The resurgence of risk assessment in criminal sentencing. *Federal Sentencing Reporter, 26*, 158-166.
- Monahan, J., & Skeem, J. (in press). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*.
- Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Roth, L., Grisso, T. & Banks, S. (2001). *Rethinking risk assessment. The MacArthur study of mental disorder and violence*. New York: Oxford.
- Morris N. (1974). *The Future of Imprisonment*. Chicago: Univ. Chicago Press
- National Conference of State Legislatures (2015). State Sentencing and Corrections Legislation. Retrieved 10/10/15 from: <http://www.ncsl.org/research/civil-and-criminal-justice/state-sentencing-and-corrections-legislation.aspx>
- Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk Assessment With Young Offenders A Meta-Analysis of Three Assessment Measures. *Criminal Justice and Behavior, 36*(4), 329-353.
- Petersilia, J. & Turner, S. (1987). *Guideline-Based Justice: The Implications for Racial Minorities*. Los Angeles, CA: RAND Corporation. Available: <http://www.rand.org/pubs/reports/R3306.html>
- Piquero, A. R., & Brame, R. W. (2008). Assessing the Race-Crime and Ethnicity-Crime Relationship in a Sample of Serious Adolescent Delinquents. *Crime & Delinquency, 54*(3), 390-422.

- Reiser, B., & Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: the normal equal variance case. *Journal of the Royal Statistical Society*, 48, 413-418.
- Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In *Handbook of Cross-Cultural Neuropsychology*, ed. E Fletcher-Janzen, T Strickland, & CR Reynolds, pp. 249--85. New York: Springer.
- Reynolds, C.R. & Suzuki, L.A. (2012). Bias in psychological assessment: An empirical review and recommendations. In *Handbook of Psychology Vol 10, Assessment Psychology*, 2nd ed., B Weiner, JR Graham, & JA Naglieri (Eds), pp. 82-113. New York: Wiley.
- Rice ME, Harris GT. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law & Human Behavior* 29: 615-620.
- Rogers, WH (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin* 13: 19–23.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: a meta-analysis. *Personnel Psychology*, 54, 297-330.
- Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual review of psychology*, 65, 693-717.
- Sackett, P.R., & Bobko, P. (July, 2015). Conceptual and Technical Issues in Conducting and Interpreting Differential Prediction Analyses. *Industrial and Organizational Psychology*, 3, 213-217.
- Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: appraising the evidence for validity and fairness. *American Psychologist*, 63(4), 215-227
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56, 302-318.
- Sentencing Project, The (2000). Reducing racial disparity in the criminal justice system: A Manual for practitioners and policymakers. Retrieved 10/10/15 from: [http://www.sentencingproject.org/doc/publications/rd\\_reducingracialdisparity.pdf](http://www.sentencingproject.org/doc/publications/rd_reducingracialdisparity.pdf)
- Sentencing Project News (July, 2015). *Risk Assessment or Race Assessment?* Retrieved 9/16/15 from: [http://www.sentencingproject.org/detail/news.cfm?news\\_id=1955](http://www.sentencingproject.org/detail/news.cfm?news_id=1955)

- Silver, E., & Miller, L. L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime & Delinquency*, 48, 138-161.
- Silver, E., Smith, W. R., & Banks, S. (2000). Constructing Actuarial Devices for Predicting Recidivism A Comparison of Methods. *Criminal Justice and Behavior*, 27(6), 733-764.
- Singh, J. P., & Fazel, S. (2010). Forensic Risk Assessment: A Metareview. *Criminal Justice and Behavior*, 37(9), 965-988.
- Skeem, J., Barnoski, R., Latessa, E., Robinson, D., & Tjaden, C. (2013). *Youth risk assessment approaches: Lessons learned and question raised by Baird et al.'s study*. Rebuttal prepared for the National Council on Crime & Delinquency (NCCD) study funded by the Office of Juvenile Justice and Delinquency Prevention (OJJDP). Retrieved 10/10/15 from: [http://risk-resilience.berkeley.edu/sites/default/files/wp-content/gallery/publications/BairdRebuttal2013\\_FINALc1.pdf](http://risk-resilience.berkeley.edu/sites/default/files/wp-content/gallery/publications/BairdRebuttal2013_FINALc1.pdf)
- Skeem, J. L., Edens, J. F., Camp, J., & Colwell, L. H. (2004). Are there ethnic differences in levels of psychopathy? A meta-analysis. *Law and Human Behavior*, 28, 505-527.
- Society for Industrial and Organizational Psychology (2003). Principles for the Validation and Use of Personnel Selection Procedures, 4<sup>th</sup> ed. Downloaded 10/10/15 from: [http://www.siop.org/\\_principles/principles.pdf](http://www.siop.org/_principles/principles.pdf)
- Starr, S.B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, 66, 803-872.
- Starr, S.B. (2015). The new profiling: Why punishing based on poverty and identity is unconstitutional and wrong. *Federal Sentencing Reporter*, 27, 229-236.
- Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, Black, and male. *Criminology*, 36, 763-797.
- Subramanian, R., Moreno, R., & Broomhead, S. (2014). *Recalibrating Justice: A Review of 2013 State Sentencing and Corrections Trends*. New York: Vera Institute of Justice <http://www.vera.org/sites/default/files/resources/downloads/state-sentencing-and-corrections-trends-2013-v2.pdf>
- Swanson, J., Swartz, M., Van Dorn, R. A., Monahan, J., McGuire, T. G., Steadman, H. J., & Robbins, P. C. (2009). Racial disparities in involuntary outpatient commitment: Are they real? *Health Affairs*, 28, 816-826.

- Tonry, M. (2012). Race, ethnicity, and punishment. In K. Reitz & J. Petersilia (Eds.), *Oxford Handbook of Sentencing and Corrections*, pp. 53-81. New York: Oxford University Press.
- Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter* 26: 167-176.
- Tonry, M., & Melewski, M. (2008). The malign effects of drug and crime control policies on Black Americans. *Crime and Justice*, 37, 1-44.
- Ulmer, J.T. (2012). Recent developments and new directions in sentencing research. *Justice Quarterly* 29: 1-40.
- Ulmer, J., Painter-Davis, N., & Tinik, L. (2014). Disproportional Imprisonment of Black and Hispanic Males: Sentencing Discretion, Processing Outcomes, and Policy Structures. *Justice Quarterly*, (ahead-of-print), 1-40.
- van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 54(2), 119-135.
- van Wingerden, S., van Wilsem, J., & Moerings, M. (2014). Pre-sentence reports and punishment: A quasi-experiment assessing the effects of risk-based pre-sentence reports on sentencing. *European Journal of Criminology*, 11, 723-744.
- Walker, S., Spohn, C., & DeLone, M. (2011). *The Color of Justice: Race, Ethnicity, and Crime in America*, 5<sup>th</sup> ed. Cengage Learning. Belmont, CA: Wadsworth.
- Walters, G. D. (2012). Psychopathy and crime: testing the incremental validity of PCL-R-measured psychopathy as a predictor of general and violent recidivism. *Law and human behavior*, 36 404-412.
- Walters, G. D., & Lowenkamp, C. T. (2015). Predicting Recidivism With the Psychological Inventory of Criminal Thinking Styles (PICTS) in Community-Supervised Male and Female Federal Offenders. *Psychological Assessment*, online first, available: <http://dx.doi.org/10.1037/pas0000210>
- Wilson, H. A., & Gutierrez, L. (2014). Does One Size Fit All? A Meta-Analysis Examining the Predictive Ability of the Level of Service Inventory (LSI) With Aboriginal Offenders. *Criminal Justice and Behavior*, 41, 196-219.
- Wroblewski, J. (2014). *2014 US Department of Justice Criminal Division Annual Letter to US Sentencing Commission*

<http://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf>

Yang, M, Wong, S.C., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin* 136: 740-767



### Endnotes

---

<sup>i</sup> Effect sizes were calculated by the first author based on data shared by Frase et al. (2015).

<sup>ii</sup> The correlation of race with age, sex, and offense type would yield imprecise estimates of race effects—and require complex interaction terms that are not compatible with the approach for testing predictive fairness. The matched sample allows specific focus on the relationship between risk and race. We report supplemental results on the eligible, non-matched sample below.

<sup>iii</sup> Because no cutoff values for small, medium, and large values of the DIF-R are available it is not possible to compare them using these benchmarks. Further, since no formulae are available to estimate the confidence intervals of the DIF-R it is not possible to determine if the DIF-R values for White and Black offenders differ significantly from one another.

<sup>iv</sup> PCRA total scores greater than 16 were recoded to 16 as only 18 offenders have a PCRA total score of 17 or 18.

<sup>v</sup> Theoretically, it is possible. Most validated risk assessment tools have predictive utilities that are essentially interchangeable (Yang, Wong & Coid, 2010). In part, this may be because a limiting process makes recidivism impossible to predict beyond a certain level of accuracy (see Monahan & Skeem, 2014). A scale can reach this limit quickly with a few maximally predictive items, before reaching a sharp point of diminishing returns. But if there is a natural limit, it can be reached via alternative routes. If measured validly, some variable risk factors (e.g., attitudes supportive of crime) predict recidivism as strongly as common risk markers (e.g., early antisocial behavior; Gendreau et al., 1996). This theoretical possibility must be balanced, however, by sobering observations about how predictive utility can be compromised when suspect risk factors are eliminated (Berk, 2009; Petersilia & Turner, 1987; Sackett et al., 2001)—particularly for short scales.

## DATA-DRIVEN DISCRIMINATION AT WORK

PAULINE T. KIM\*

### ABSTRACT

*A data revolution is transforming the workplace. Employers are increasingly relying on algorithms to decide who gets interviewed, hired, or promoted. Although data algorithms can help to avoid biased human decision-making, they also risk introducing new sources of bias. Algorithms built on inaccurate, biased, or unrepresentative data can produce outcomes biased along lines of race, sex, or other protected characteristics. Data mining techniques may cause employment decisions to be based on correlations rather than causal relationships; they may obscure the basis on which employment decisions are made; and they may further exacerbate inequality because error detection is limited and feedback effects compound the bias. Given these risks, I argue for a legal response to classification bias—a term that describes the use of classification schemes, such as data algorithms, to sort or score workers in ways that worsen inequality or disadvantage along the lines of race, sex, or other protected characteristics.*

*Addressing classification bias requires fundamentally rethinking antidiscrimination doctrine. When decision-making algorithms produce biased outcomes, they may seem to resemble familiar disparate*

---

\* Daniel Noyes Kirby Professor of Law, Washington University School of Law, St. Louis, Missouri. This Article benefitted from the comments of participants at workshops at Washington University School of Law, the University of Denver Sturm College of Law, and the 2015 Colloquium on Scholarship in Employment and Labor Law. I would like to thank Scott Baker, Marion Crain, Lee Epstein, Peggie Smith, David Law, and Neil Richards for their helpful comments. I am also indebted to participants at the June 2016 Privacy Law Scholars' Conference—especially danah boyd, Solon Barocas, Andrew Selbst, and Mark MacCarthy—for their valuable feedback, and to Erika Hanson and Jae Ryu for outstanding research assistance.

*impact cases; however, mechanical application of existing doctrine will fail to address the real sources of bias when discrimination is data-driven. A close reading of the statutory text suggests that Title VII directly prohibits classification bias. Framing the problem in terms of classification bias leads to some quite different conclusions about how to apply the antidiscrimination norm to algorithms, suggesting both the possibilities and limits of Title VII's liability-focused model.*

## TABLE OF CONTENTS

INTRODUCTION . . . . .	860
I. THE IMPACT OF DATA ANALYTICS ON WORKPLACE EQUALITY. . . . .	869
A. <i>The Promise of Workforce Analytics</i> . . . . .	869
B. <i>The Risks of Workforce Analytics</i> . . . . .	874
C. <i>Types of Harm</i> . . . . .	883
1. <i>Intentional Discrimination</i> . . . . .	884
2. <i>Record Errors</i> . . . . .	885
3. <i>Statistical Bias</i> . . . . .	886
4. <i>Structural Disadvantage</i> . . . . .	888
D. <i>Classification Bias</i> . . . . .	890
II. ALTERNATIVE SYSTEMS OF REGULATION . . . . .	892
A. <i>The Market Response</i> . . . . .	892
B. <i>Privacy Rights</i> . . . . .	897
III. THE ANTIDISCRIMINATION RESPONSE . . . . .	901
A. <i>The Conventional Account of Title VII</i> . . . . .	902
B. <i>A Closer Reading</i> . . . . .	909
C. <i>Addressing Classification Bias</i> . . . . .	916
1. <i>Data on Protected Class Characteristics</i> . . . . .	917
2. <i>Relevant Labor Market Statistics</i> . . . . .	918
3. <i>Employer Justifications</i> . . . . .	920
4. <i>The Bottom-Line Defense</i> . . . . .	923
D. <i>A Note on Ricci v. DeStefano</i> . . . . .	925
E. <i>The Limits of the Liability Model</i> . . . . .	932
CONCLUSION. . . . .	936

## INTRODUCTION

The data revolution has come to the workplace. Just as the analysis of large datasets has transformed the businesses of baseball, advertising, medical care, and policing, it is radically altering how employers manage their workforces. Employers are increasingly relying on data analytic tools to make personnel decisions, thereby affecting who gets interviewed, hired, or promoted.<sup>1</sup> Using highly granular data about workers' behavior both on and off the job, entrepreneurs are building models that they claim can predict future job performance.<sup>2</sup> Sometimes called workforce or people analytics, these technologies aim to help employers recruit talented workers, screen for eligible candidates in an applicant pool, and predict an individual's likelihood of success at a particular job.<sup>3</sup>

Proponents of the new data science claim that it will not only help employers make better decisions faster, but that it is fairer as well because it can replace biased human decision makers with "neutral" data.<sup>4</sup> However, as many scholars have pointed out, data are not neutral, and algorithms can discriminate.<sup>5</sup> Large datasets often

---

1. See, e.g., George Anders, *Who Should You Hire? LinkedIn Says: Try Our Algorithm*, FORBES (Apr. 10, 2013, 4:31 PM), <http://www.forbes.com/sites/georgeanders/2013/04/10/who-should-you-hire-linkedin-says-try-our-algorithm> [<https://perma.cc/M7NF-SJJD>]; Jeanne Meister, *2014: The Year Social HR Matters*, FORBES (Jan. 6, 2014, 10:21 AM), <http://www.forbes.com/sites/jeannemeister/2014/01/06/2014-the-year-social-hr-matters/> [<https://perma.cc/L6SJ-VMJE>]; Claire Cain Miller, *Can an Algorithm Hire Better Than a Human?*, N.Y. TIMES: THEUPSHOT (June 25, 2015), <http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html> [<https://perma.cc/PKM6-4JY4>].

2. See, e.g., Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), <http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html> [<https://perma.cc/3X99-EM4X>].

3. See Josh Bersin, *Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age*, FORBES (Feb. 17, 2013, 8:00 PM), <http://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/> [<https://perma.cc/W69F-3BAM>].

4. See, e.g., *id.* (discussing workforce analytics as the superior alternative to employment decisions "made on gut feel"); Lohr, *supra* note 2 (examining views of many proponents of workforce analytics).

5. See, e.g., Solon Barocas & Andrew D. Selbst, Essay, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674 (2016); danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO. COMM. & SOC'Y 662, 666-68 (2012); Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 35 (2013); Joshua A. Kroll, Joanna Huey, Solon Barocas,

contain errors in individual records, and these errors may not be randomly distributed. Algorithms that are built on inaccurate, biased, or unrepresentative data can in turn produce outcomes biased along lines of race, sex, or other protected characteristics. When these automated decisions are used to control access to employment opportunities, the results may look very similar to the systematic patterns of disadvantage that motivated antidiscrimination laws. What is novel is that the discriminatory effects are data-driven.

Of course, employers have always done things such as recruiting, hiring, evaluating, promoting, and terminating employees, but data models do not rely on traditional indicia like formal education or on-the-job experience. Instead, they exploit the information in large datasets containing thousands of bits of information about individual attributes and behaviors. Third-party aggregators harvest information from the internet about job applicants, including detailed information about their social networking habits—how many contacts they have, who those contacts are, how often they post messages, who follows them, and what they like.<sup>6</sup> Similarly, monitoring devices collect data on the workplace behaviors of current employees, recording information such as where they go during the day, how often they speak with others and for how long, and who initiates the conversation and who terminates it.<sup>7</sup> Employers can also obtain information about their employees' off-duty behavior. As employees spend more of their personal time online, third parties can collect information on those activities, aggregate it with other data, and share it with employers.<sup>8</sup> Growing participation in wellness programs means that employees increasingly share

---

Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. (forthcoming 2017) (manuscript at 29-35), <https://ssrn.com/abstract=2765268> [<https://perma.cc/CL85-DUKK>]; Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013), <http://foreignpolicy.com/2013/05/10/think-again-big-data/> [<https://perma.cc/V9XM-MNJ6>].

6. See Michael Fertik, *Your Future Employer Is Watching You Online. You Should Be, Too.*, HARV. BUS. REV. (Apr. 3, 2012), <https://hbr.org/2012/04/your-future-employer-is-watchi> [<https://perma.cc/XZ58-D5DC>]; Meister, *supra* note 1.

7. See Don Peck, *They're Watching You at Work*, ATLANTIC (Dec. 2013), <https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/> [<https://perma.cc/92WJ-6VUD>].

8. See, e.g., Esther Kaplan, *The Spy Who Fired Me: The Human Costs of Workplace Monitoring*, HARPER'S MAG., Mar. 2015, at 31-32, 35.

information about their offline behaviors as well, reporting such things as how often they exercise or what they eat.<sup>9</sup> Data miners use this information to make health-related predictions, such as whether an employee is pregnant or trying to conceive.<sup>10</sup> Aggregating these various data sources can produce a rich and highly detailed profile of individual workers.<sup>11</sup>

This volume of information requires some form of automatic processing. No human brain can keep in view all of the thousands of data points about an individual. And so, algorithms are developed to make sense of it all—to screen, score, and evaluate individual workers for particular jobs. These algorithms are the tools of workforce analytics. For example, a company called Gild offers a “smart hiring platform” to help companies find “the right talent quicker.”<sup>12</sup> Gild uses an algorithm that

crunches thousands of bits of information in calculating around 300 larger variables about an individual: the sites where a person hangs out; the types of language, positive or negative, that he or she uses to describe technology of various kinds; self-reported skills on LinkedIn; [and] the projects a person has worked on, and for how long

as well as traditional criteria such as education and college major.<sup>13</sup> Other services screen large pools of applicants, automating the

---

9. See generally Jay Hancock, *Workplace Wellness Programs Put Employee Privacy at Risk*, CNN (Oct. 2, 2015, 12:37 PM), <http://www.cnn.com/2015/09/28/health/workplace-wellness-privacy-risk-exclusive/> [https://perma.cc/X9RY-X4VZ].

10. See Valentina Zarya, *Employers Are Quietly Using Big Data to Track Employee Pregnancies*, FORTUNE (Feb. 17, 2016, 5:36 PM), <http://fortune.com/2016/02/17/castlight-pregnancy-data/> [https://perma.cc/MA3W-DDZQ].

11. See, e.g., Peck, *supra* note 7; Sanjeev & Sandeep Sardana, *Big Data: It's Not a Buzzword, It's a Movement*, FORBES (Nov. 20, 2013, 12:05 PM), <http://www.forbes.com/sites/sanjeivsardana/2013/11/20/bigdata/> [https://perma.cc/PU97-VFA9].

12. *Our Story*, GILD, <https://www.gild.com/company> [https://perma.cc/Q8QF-RPGB]; see Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. TIMES (Apr. 27, 2013), <http://www.nytimes.com/2013/04/28/technology/how-big-data-is-playing-recruiter-for-specialized-workers.html> [https://perma.cc/XAF6-SKXC].

13. Richtel, *supra* note 12; see also Vivian Giang, *Why New Hiring Algorithms Are More Efficient—Even If They Filter Out Qualified Candidates*, BUS. INSIDER (Oct. 25, 2013, 10:51 AM), <http://www.businessinsider.com/why-its-ok-that-employers-filter-out-qualified-candidates-2013-10> [https://perma.cc/3XLE-GH6V] (describing how Bright.com uses “data and algorithms to match candidates up with potential jobs and hiring managers with star performers”).

process of selecting the most promising candidates for employers.<sup>14</sup> One company examines hundreds of variables about job seekers, analyzes a firm's past hiring practices, and then recommends only those applicants it believes the employer will be interested in hiring. Other firms are developing computer games that record thousands of data points about how individuals play, such as what decisions they make and how long they hesitate before deciding, in order to uncover patterns that can identify successful employees.<sup>15</sup> Employers can then use these tools to make hiring or promotion decisions.

The actual impact on employment opportunities is difficult to document because information about how developers construct these algorithms is considered proprietary, and personnel data is confidential. Nevertheless, some publicly available examples suggest there is reason for concern. One company seeking to identify which employees would stay longer found that the distance between home and the workplace is a strong predictor of job tenure.<sup>16</sup> If a hiring algorithm relied on that factor, it would likely have a racially disproportionate impact, given that discrimination has shaped residential patterns in many cities. Other studies involving internet advertising illustrate how algorithms that learn from behavioral patterns can discriminate. For example, Latanya Sweeney has shown that Google searches for African American-associated names produce more advertisements for criminal background checks than searches for Caucasian-associated names, likely reflecting past patterns in users' search behavior.<sup>17</sup> Amit Datta, Michael Carl Tschantz, and Anupam Datta have demonstrated gender differences in the delivery of online ads to jobseekers, with identified male users "receiv[ing] more ads for a career coaching service that promoted high pay jobs," while female users received more generic ads.<sup>18</sup> Similarly, a field study by Anja Lambrecht and Catherine

---

14. See Miller, *supra* note 1.

15. See, e.g., Peck, *supra* note 7.

16. See Dustin Volz, *Silicon Valley Thinks It Has the Answer to Its Diversity Problem*, ATLANTIC (Sept. 26, 2014), <http://www.theatlantic.com/politics/archive/2014/09/silicon-valley-thinks-it-has-the-answer-to-its-diversity-problem/431334/> [<https://perma.cc/VA6N-6W53>].

17. See Latanya Sweeney, *Discrimination in Online Ad Delivery*, COMM. ACM, May 2013, at 44, 46-47.

18. See Amit Datta, Michael Carl Tschantz & Anupam Datta, *Automated Experiments on Ad Privacy Settings*, PROC. ON PRIVACY ENHANCING TECHS., Apr. 2015, at 92, 92-93; see also Amit Datta, Anupam Datta, Deirdre K. Mulligan & Michael Carl Tschantz, *Discrimination*



Tucker revealed that an internet ad for STEM (science, technology, engineering and math) jobs was far less likely to be shown to women than men.<sup>19</sup> These examples did not necessarily result from intentional bias, but the discriminatory effects were nevertheless real.

While workforce analytics are transforming employers' personnel practices, the legal world has only just begun to take notice. Privacy law scholars have raised concerns about the growth of big data, asking what limits the law should place on the collection of particularly sensitive personal information, or whether it should regulate "data flows" or downstream uses of this information.<sup>20</sup> Although much of the focus has been on problems caused by inaccurate data records or unexpected and invasive uses of sensitive personal information,<sup>21</sup> these scholars have also sounded alarms that big data may produce biased outcomes. Of the handful of commenters who have addressed the employment context, most have simply raised questions about the discriminatory potential of data analytics,<sup>22</sup> without deeply theorizing the nature of the harms that these technologies threaten for workers. And to the extent that legal scholars have considered how the law might respond, they have confined their analysis to narrowly applying existing doctrine.<sup>23</sup>

---

in Online Personalization: A Multidisciplinary Inquiry 3-5 (Mar. 13, 2016) (unpublished manuscript) (on file with author) (describing experiment and analyzing possible legal response).

19. See Anja Lambrecht & Catherine Tucker, Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads 2, 10-12 (Oct. 13, 2016) (unpublished manuscript), <https://ssrn.com/abstract=2852260> [<https://perma.cc/3PGF-CVTW>].

20. See, e.g., Danielle Keats Citron & Frank Pasquale, Essay, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4, 7-8, 18-22 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 94-96, 98-99, 101, 103-09, 123-27 (2014). See generally Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 409 (2014); Neil M. Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 20 STAN. TECH. L. REV. (forthcoming 2017), <https://ssrn.com/abstract=2655719> [<https://perma.cc/58A8-SCZB>].

21. See Citron & Pasquale, *supra* note 20, at 4; Crawford & Schultz, *supra* note 20, at 96-99.

22. See, e.g., EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51-53 (2014), <https://perma.cc/LE9N-PA9D>; Citron & Pasquale, *supra* note 20, at 4; danah boyd & Kate Crawford, Six Provocations for Big Data (Sept. 21, 2011) (unpublished manuscript), <https://ssrn.com/abstract=1926431> [<https://perma.cc/3JUG-FAFJ>]; Alex Rosenblat, Kate Wikelius, danah boyd, Seeta Peña Gangadhoran & Corrine Yu, Data & Civil Rights: Employment Primer (Oct. 30, 2014) (unpublished manuscript), <https://ssrn.com/abstract=2541512> [<https://perma.cc/398Q-F5MZ>].

23. See, e.g., Barocas & Selbst, *supra* note 5, at 694-712 (applying existing Title VII doc-

Workforce analytics pose an entirely new set of challenges to equality that calls for fundamentally rethinking antidiscrimination doctrine. Proponents of workforce analytics argue that data models can avoid reliance on biased human decision-making.<sup>24</sup> Skeptics warn that data is not neutral and that workforce analytics threaten to introduce new forms of bias or exacerbate existing ones.<sup>25</sup> But there is a third possibility as well—employers and researchers can use data to diagnose where and how cognitive or structural biases are currently operating in ways harmful to disadvantaged groups. Thus, the impact of workforce analytics will depend to a large extent on the choices that are made about how to deploy these technologies. And those choices will be shaped in turn by the legal environment in which firms operate.

The harms threatened by biased algorithms are not easily captured by traditional antidiscrimination law, which tends to focus on a specific “bad actor” and individual victims. Of course, a prejudiced employer might hide its discriminatory intent behind a biased data model. Such a scenario poses no particular conceptual challenge, although proof may be difficult as a practical matter. Even without any deliberate intent, a model may be biased in the statistical sense. Choices in the coding of information, errors in the data, reliance on unrepresentative samples, or the selection of variables for exclusion or inclusion might produce a model that is inaccurate in a systematic way.<sup>26</sup> When those systematic errors coincide with protected class status and operate to reduce opportunities for already disadvantaged groups, it should trigger the same concerns about workplace equality that motivated antidiscrimination laws.

The nature of algorithmic decision-making raises particular concern when employers rely on these models to make personnel decisions. Data mining techniques used to build the algorithms seek to uncover any statistical relationship between variables present in the data, regardless of whether the reasons for the relationship are understood. As a result, if employers rely on these models, they may deny employees opportunities based on unexplained correlations and make decisions that turn on factors with no clear causal

---

trine).

24. See *supra* note 4 and accompanying text.

25. See *supra* note 5.

26. See *infra* Part I.B.

connection to effective job performance. Because of limited opportunities for error correction, and the possibility of reinforcing feedback effects, these models may not only introduce but actually worsen bias and inequality. Given these risks, the law ought to be concerned with what I call “classification bias.” Classification bias occurs when employers rely on classification schemes, such as data algorithms, to sort or score workers in ways that worsen inequality or disadvantage along the lines of race, sex, or other protected characteristics.

Classification bias may seem amenable to challenge under disparate impact doctrine, which targets facially neutral employment practices that have disparate effects on racial minorities or other protected classes.<sup>27</sup> However, a mechanical application of existing disparate impact doctrine will fail to meet the particular risks that workforce analytics pose. That doctrine evolved to address employer use of tests purporting to measure workers’ abilities, and therefore focused on the validity of those measures and their relevance to a particular job.<sup>28</sup> In contrast, data mining models do not rest on psychological or any other theories of human behavior. Instead, these models simply mine the available data, looking for statistical correlations that connect seemingly unrelated variables, such as patterns of social media behavior, with workplace performance.<sup>29</sup> As a result, they pose a different set of risks—risks that existing doctrine does not address well.

As an example, disparate impact doctrine provides a defense if an employer can show that a test is “job related ... and consistent with business necessity.”<sup>30</sup> In the case of workforce analytics, the data algorithm by definition relies on variables that are correlated in some sense with the job. So to ask whether the model is “job related” in the sense of “statistically correlated” is tautological. The more important question in the context of data mining is what does the correlation mean? Is the statistical relationship it uncovers causal, such that it provides a reliable basis for predicting future behavior?

---

27. See Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(2) (2012); *Griggs v. Duke Power Co.*, 401 U.S. 424, 430-31 (1971).

28. See Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 755-60 (2006).

29. See Fertik, *supra* note 6.

30. See Civil Rights Act of 1991 § 105, 42 U.S.C. § 2000e-2(k)(1)(A)(i).

Or does it result from erroneously coded information, an unrepresentative sample, omitted variable bias, or some other data problems? Because the risks to workplace equality posed by data mining algorithms arise from different sources, existing disparate impact doctrine will not be adequate to address the risks they pose.

Addressing the possibilities and risks of data analytics for workplace equality requires taking a fresh look at antidiscrimination law, unencumbered by the specific doctrinal details that have developed under Title VII. Revisiting the statutory text suggests that Title VII directly prohibits classification bias. More specifically, section 703(a)(2) forbids employer practices that “classify” employees or applicants “in any way which would deprive or tend to deprive” them of employment opportunities because of protected class characteristics.<sup>31</sup> By focusing on the consequences of employers’ classification schemes, this reading offers a more relevant frame for addressing the challenges that workforce analytics pose.

Thinking about the problem in terms of classification bias leads to some quite different conclusions about how the antidiscrimination norm should apply to data models.<sup>32</sup> For example, if the goal is to discourage classification bias, then the law should not forbid the inclusion of race, sex, or other sensitive information as variables, but seek to preserve these variables, and perhaps even include them in some complex models.<sup>33</sup> Similarly, this perspective suggests that those who use data mining models should bear the burden of demonstrating the accuracy and representativeness of the data used to construct the models, rather than requiring complainants to identify the flaws giving rise to biased outcomes.<sup>34</sup>

Addressing the challenges of workforce analytics using a theory of classification bias also reveals the limitations of the backward-looking, liability-focused model of legal regulation embodied by Title

---

31. The full text of subsection (a)(2) reads:

It shall be an unlawful employment practice for an employer ... to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual’s race, color, religion, sex, or national origin.

Civil Rights Act of 1964 § 703(a)(2), 42 U.S.C. § 2000e-2(a)(2).

32. See *infra* Part III.C.

33. See *infra* Part III.C.1.

34. See *infra* Part III.C.3.

VII.<sup>35</sup> Because of the diffuse nature of the harms and the significant resources that would be required to challenge biased algorithms, it may be difficult to incentivize individual plaintiffs to enforce a prohibition on classification bias. Even more problematic, a strong liability regime intended to address the use of biased algorithms may discourage employers from trying to understand whether these tools have disparate effects or may discourage them from using algorithms at all. If the law swings too far in this direction, it would avoid the costs of biased algorithms but also eliminate any potential positive effects that data analytics might have on diagnosing and counteracting cognitive and structural biases already affecting workplaces. Resolving this dilemma may require looking beyond liability-focused legal models to alternatives such as *ex ante* regulation, licensing models, or the development of technological solutions.

In considering the impact of data analytics on workplace equality and the appropriate legal response, this Article proceeds as follows. Part I surveys the psychological and structural factors that contribute to bias in the contemporary workplace and considers the potential for data models to eliminate that bias. Replacing human decision makers with a computer algorithm may prevent certain types of cognitive biases from operating but is unlikely to reach other types of structural disadvantage that may result from the way work is organized. At the same time, widespread reliance on decision-making algorithms risks introducing new forms of bias or exacerbating existing ones. Part I surveys those risks, catalogues the types of harm that may result from reliance on algorithms in the workplace, and then argues for recognizing classification bias as a distinct type of threat to workplace equality.

Part II considers whether other responses—aside from antidiscrimination law—can effectively address classification bias and concludes that neither market forces nor traditional forms of privacy protection are likely to be successful. The nature of labor markets are such that employers will not reliably receive signals if their employment practices produce bias against minority groups. And privacy protections typically focus on individual harms rather than addressing the group-based disadvantages that are the principal concern of antidiscrimination law.

---

35. See *infra* Part III.E.

In Part III, I consider the limits and possibilities of existing antidiscrimination law. Mechanical application of existing Title VII doctrine is unlikely to be successful in addressing the equality challenges that workforce analytics pose. Neither disparate treatment nor current disparate impact doctrine completely captures the types of risks threatened by data models. Instead, antidiscrimination law should be adapted to meet these unique risks. Part III argues that a close reading of the statutory text shows that Title VII *does* prohibit classification bias, and considers what a robust response to this form of discrimination should look like.

More specifically, it argues that an effective legal response will depart from traditional disparate impact doctrine in several ways. For example, employers should not be able to justify reliance on a biased model merely by showing a statistical relationship but should bear the burden of showing that the model is statistically valid and substantively meaningful. At the same time, an employer should be permitted to rely on a “bottom-line” defense if its use of a model as part of a larger selection process does not produce discriminatory results.

After considering how the law should respond, Part III briefly explains why the Supreme Court’s decision in *Ricci v. DeStefano* poses no obstacle to enforcing a prohibition on classification bias. Finally, it considers the limitations of classification bias theory and suggests some alternatives to a liability-based regime.

## I. THE IMPACT OF DATA ANALYTICS ON WORKPLACE EQUALITY

### A. *The Promise of Workforce Analytics*

The use of data analytics offers the potential to *reduce* bias in employment. Proponents of the technology claim that algorithms do just that by eliminating the subjective biases and personal predilections of a human resources manager. For example, the goal of Gild is “to build machines that ... eliminate human bias.”<sup>36</sup> Pointing to the many ways in which human decision-making is biased, these services offer to find overlooked talent that better matches a company’s needs and, in turn, to produce a more diverse workforce.

---

36. See Richtel, *supra* note 12.

These claims are consistent with scholarly accounts of how human bias distorts personnel decisions, even in the absence of a conscious discriminatory motive.<sup>37</sup> Charles Lawrence argues that unconscious prejudices may lead to discrimination even when the decision maker is unaware of, and would disclaim, any prejudicial intent.<sup>38</sup> Similarly, Linda Krieger and other scholars explain how ordinary cognitive processes naturally lead people to create mental categories.<sup>39</sup> When these categories coincide with race or gender differences, they can distort the perceptions of supervisors and managers in ways that tend to confirm societal biases. More recently, a great deal of attention has focused on implicit bias.<sup>40</sup> Scholars point to the results of the Implicit Associations Test to argue that people typically associate negative characteristics more strongly with disfavored groups.<sup>41</sup> These negative associations can result in adverse decisions for members of those groups, even when the decision maker intends to act fairly and believes that she is doing so.<sup>42</sup>

Although these theories differ as to the precise mechanism at work, they are alike in pointing to processes that occur outside of conscious awareness. They suggest that automatic processes—the ways in which our brains naturally function—can produce biased judgments. As a result, these effects are not readily visible to the decision maker, even upon self-reflection.<sup>43</sup> Individuals who strongly embrace nondiscrimination and equality norms may be particularly

---

37. See, e.g., Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 322 (1987).

38. See *id.*

39. See, e.g., Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1186-87 (1995).

40. See, e.g., R. Richard Banks, Jennifer L. Eberhardt & Lee Ross, *Discrimination and Implicit Bias in a Racially Unequal Society*, 94 CALIF. L. REV. 1169 (2006); Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CALIF. L. REV. 945 (2006); Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CALIF. L. REV. 969 (2006); Jerry Kang, *Rethinking Intent and Impact: Some Behavioral Realism About Equal Protection*, 66 ALA. L. REV. 627 (2015); Jerry Kang & Kristin Lane, *Seeing Through Colorblindness: Implicit Bias and the Law*, 58 UCLA L. REV. 465 (2010).

41. For reviews of the social science literature on implicit bias, see Greenwald & Krieger, *supra* note 40, at 951-58; Kang & Lane, *supra* note 40, at 473-81.

42. See, e.g., Kang & Lane, *supra* note 40, at 468-89.

43. See Lawrence, *supra* note 37, at 336-39; see also Krieger, *supra* note 39, at 1217.

resistant to recognizing the operation of bias in their mental processing because of the cognitive dissonance that would result.<sup>44</sup>

The claim of workforce analytics is that algorithms can replace fallible human judgments with neutral, unbiased data to improve decision-making.<sup>45</sup> The chief scientist at Gild put it this way: “Let’s put everything in and let the data speak for itself.”<sup>46</sup> The proponents of data science are right to point out that traditional employment practices—relying as they often do on subjective assessments, intuition, and limited human cognition—may entail considerable amounts of bias. However, as discussed in Part I.B below, algorithms are not always neutral either. Depending on the choices made in collecting and coding information and building models, data analytics risk replicating existing biases or introducing new ones.<sup>47</sup> So although algorithms offer the potential for avoiding or minimizing bias, the real question is how the biases they may introduce compare with the human biases they avoid.

Whatever their promise for eliminating cognitive biases, algorithms will not counteract structural forms of workplace bias. This type of bias results not from cognitive processes but from structural forces that shape opportunities differently for different types of people. Numerous scholars have argued that workplaces are often organized in ways that systematically disadvantage women or minorities.<sup>48</sup> For example, when training and advancement opportunities are informally distributed in a firm through social networks, women or racial minorities who have less extensive networks may be disadvantaged. Similarly, work that requires long hours or unpredictable schedules may place particular burdens on women, who are often the primary caretakers of their children. These types of choices about workplace organization may not reflect intent to exclude, and, therefore, like the cognitive processes described above,

---

44. See Lawrence, *supra* note 37, at 337.

45. See Richtel, *supra* note 12.

46. *Id.* (quoting Vivienne Ming, chief scientist at Gild).

47. See *supra* note 5.

48. See, e.g., Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 11 (2006); Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. C.R.-C.L. L. REV. 91, 104 (2003); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 468-74 (2001).



their impact on disadvantaged groups is not readily visible to managers.<sup>49</sup>

Relying on data models instead of human decision-making is unlikely to counter structural forms of bias because these models take existing workplace structures as givens. For example, if reduced access to social networks in a firm hampers minority employees' chances of promotion, relying on data to make those promotion decisions will not remedy the fact that minority employees are receiving less mentoring and training. Similarly, data-driven hiring decisions will not alter the reality that unpredictable work schedules will take a greater toll on workers with caregiving responsibilities, who are more often women. Thus, merely relying on data analytics instead of human judgments will not address forms of disadvantage that result from biased workplace structures.

On the other hand, data can be a useful tool for *diagnosing* both cognitive and structural forms of bias. Rather than using workforce analytics to *make* decisions, firms could deploy close analysis of employment-related data to assess the decision-making process itself, thereby uncovering hidden biases and prompting efforts to counteract them. One service, Textio, used language analysis to determine that certain phrases in job postings—for example, military analogies like “mission critical”—appear to reduce the proportion of women who apply.<sup>50</sup> Employers committed to recruiting a diverse workforce might learn how to craft language likely to attract a more diverse applicant pool from such a program. Cognitive science teaches that individuals tend to remember facts that confirm their preexisting beliefs about the world.<sup>51</sup> Krieger and others explain how this phenomenon might lead supervisors to remember negative information about members of disfavored groups but to disregard similar information about in-group members.<sup>52</sup> Data could be a useful corrective to such biased perceptions, highlighting for managers when their recall about particular workers may be faulty.

Supervisors who are not themselves biased might nevertheless fail to recognize how earlier discriminatory decisions continue to shape current outcomes. An initial discriminatory decision that

---

49. *See id.* at 470-71.

50. *See* Miller, *supra* note 1.

51. *See* Krieger, *supra* note 39, at 1203.

52. *See id.* at 1209.

created a pay differential between men and women can have effects years later, even if every subsequent decision regarding individual raises is entirely fair and neutral.<sup>53</sup> A current supervisor, having directly observed only unbiased decisions in recent years, might view the differential in wages as justified. By decomposing the factors contributing to current salary or by comparing salary to discrete measures of productivity, data analysis might make visible the current effects of past discrimination, rather than allowing those outcomes to appear natural and inevitable.

Employers can also use data to identify sources of structural bias that disadvantage certain groups. In the example cited in the introduction, Evolv, the company that identified the distance between home and the workplace as a predictor of employee job tenure, decided not to use this factor in its hiring algorithm because it understood that housing patterns are correlated with race and that relying on that correlation might result in discrimination.<sup>54</sup> In addition to eliminating the factor as a basis for decision-making, an employer might use the information to examine whether its workplace practices make it more difficult for employees who travel long distances to succeed. A firm committed to a diverse workforce but located in a city with a segregated housing market might consider policies like flex-time or benefits like public transit passes in order to relieve a commuting burden that falls more heavily on already disadvantaged groups.

Data analytics thus hold the potential to reduce biases and increase opportunities in the workplace for traditionally disadvantaged groups. But much depends on how data are used. When employers use analytics to evaluate personnel policies and procedures, data can help to diagnose where workplace structures or

---

53. Consider, for example, the facts in *Ledbetter v. Goodyear Tire & Rubber Co.*, 550 U.S. 618, 621-22 (2007). The plaintiff in that case, Lilly Ledbetter, worked for Goodyear Tire for nearly twenty years. *Id.* at 621. She alleged that several supervisors had given her poor evaluations because of her sex and that those discriminatory evaluations continued to result in her receiving lower pay than her male counterparts throughout her employment with the defendant. *Id.* at 622. The Supreme Court dismissed her claims on the grounds that no discriminatory pay decisions had been made during the statutory “charging period”—the last 180 days before she filed with the Equal Employment Opportunity Commission. *Id.* at 624-32. Congress eventually overturned the decision in the Lily Ledbetter Fair Pay Act of 2009, Pub. L. No. 111-2, § 3, 123 Stat. 5, 5-6 (2009) (amending 42 U.S.C. § 2000e-5(e)).

54. See Volz, *supra* note 16.

organizations inadvertently disadvantage or exclude members of certain groups. Relying on data analytics to sort applicants and employees may also reduce bias if these models are less biased than the subjective human decision makers they replace. Whether that is the case, however, depends a great deal on how the algorithms are constructed and deployed. As the next Section explores, there are numerous reasons to be concerned that workplace analytics may introduce bias or worsen existing patterns of disadvantage.

### *B. The Risks of Workforce Analytics*

Although data analytic tools offer the potential for countering biased decision-making processes and workplace structures, these same tools also risk reinforcing existing discrimination or introducing new forms of bias. Employers have long used data to sort and rank workers—for example, through preemployment tests, psychological screens, or productivity requirements. These traditional uses of data metrics to measure and evaluate can raise concerns about bias, and they have faced legal challenges.<sup>55</sup> However, the new workforce science poses distinct risks. With traditional forms of testing, employers generally started by identifying skills or attributes thought relevant to job performance and then relied on test professionals to develop measures of those skills or attributes. These forms of testing collected limited amounts of targeted information about applicants or employees. In contrast, data models today take advantage of the vastly greater quantity of data available and mine it to discover novel correlations. That data may contain information about attributes or behaviors, such as social media usage, that have no clear connection with job performance.

In order to build a model, its creators must select the data that they will use to build it—the “training data.”<sup>56</sup> The actual data mining occurs when the data are analyzed using statistical techniques

---

55. See, e.g., *Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975); *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).

56. See Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, in *DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY: DATA MINING AND PROFILING IN LARGE DATABASES* 3, 3-4 (Bart Custers, Toon Calders, Bart Schermer & Tal Zarsky eds., 2013).

to uncover patterns.<sup>57</sup> The data miner is not testing any particular hypotheses or explanations; instead, the process reveals statistical relationships among variables present in the data.<sup>58</sup> What the data miner finds thus depends on the data examined. The correlations may be causal or the relationship may be entirely coincidental.<sup>59</sup> Data mining is generally unconcerned with the reasons for the correlation.<sup>60</sup> So long as the relationships discovered are thought to be robust, the data model may use them to classify or predict future cases.<sup>61</sup> So, for example, a data model might find that individuals who “like” certain items on Facebook have higher intelligence.<sup>62</sup> Data mining cannot explain this relationship, but a model may nevertheless predict that applicants who share that characteristic are better workers and recommend their selection over those who do not.

In their article *Big Data’s Disparate Impact*, Solon Barocas and Andrew Selbst provide a taxonomy of ways that the data mining process can result in adverse impact on protected groups.<sup>63</sup> One of the first steps in building a model is identifying the target variable—in other words, defining the outcome of interest<sup>64</sup>—and defining which outcomes are desired by categorizing them.<sup>65</sup> Doing so in the employment context is not simple. Unlike credit card charges, which can be categorized with complete certainty as fraudulent or not, the category of “good employee” is not self-

---

57. See *id.* at 9 (“[T]he ... data-mining stage ... [occurs when] the data are analyzed in order to find patterns or relations. This is done using mathematical algorithms.”).

58. See *id.* at 7 (explaining that data mining differs from traditional statistical analysis, which begins with a hypothesis, because data mining generates hypotheses from the data itself).

59. See *id.* at 16-17.

60. See *id.*

61. See *id.* at 16.

62. See, e.g., Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. NAT’L ACAD. SCI. U.S. 5802, 5805 (2013) (showing that records of an individual’s Facebook “likes” can be used to accurately predict personal characteristics such as race, gender, sexual orientation, religious and political views, and intelligence); see also Toon Calders & Indrè Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 43, 45-47.

63. See Barocas & Selbst, *supra* note 5, at 677-93.

64. See *id.* at 678.

65. This process is referred to as defining “class labels” for the target variable. See *id.* at 678-79.

evident.<sup>66</sup> In order to build a model, the meaning of “good employee” must be specified in a way that the machine can understand, namely “in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example.”<sup>67</sup> Using a more holistic definition of “good” would require someone to create a measure that captures that quality and to apply it to particular individuals in order for the machine to know what it is looking for in future cases.<sup>68</sup>

As Barocas and Selbst explain, this process of classifying individuals risks reintroducing the human biases the data analysts are seeking to avoid.<sup>69</sup> If the data miner chooses to rely on only “objective” measures for the target variable, this will introduce bias of a different kind, by valuing quantifiable measures of performance over softer skills like leadership or collaboration. In order to build a predictive model, the data miner must label and classify the training data—a “necessarily subjective process of translation”<sup>70</sup>—and these choices may introduce biases against protected groups.<sup>71</sup>

The selection of the training data will affect the outcome of the model as well. As Barocas and Selbst explain, “what a model learns depends on the examples to which it has been exposed.”<sup>72</sup> The training data may incorporate biased judgments, as, for example, when they include supervisors’ evaluations or previous hiring decisions that were colored by prejudice or distorted by cognitive bias.<sup>73</sup> Because the model will accept those characterizations “as ground truth,”<sup>74</sup> it will inevitably reflect those biases in the outcomes it produces. Factual errors may exist in the data as well, and those errors may be more frequent for members of certain groups, rendering the model less accurate when applied to members of those groups.<sup>75</sup> Another concern is that the data may be unrepresentative in that different groups are not represented in proportion to their

---

66. *See id.* at 679.

67. *Id.*

68. *See id.*

69. *Id.* at 680.

70. *Id.* at 678.

71. *See id.*

72. *Id.* at 680.

73. *See id.* at 682.

74. *Id.*

75. *See id.* at 684; EXEC. OFFICE OF THE PRESIDENT, *supra* note 22, at 52.

presence in the population.<sup>76</sup> Big datasets, which often supply the training data for workforce analytics, are more likely to exclude members of minority groups and disadvantaged populations, those “who live on big data’s margins ... and whose lives are less ‘datafied’ than the general population’s.”<sup>77</sup> If the data collection process systematically captures less information about certain groups, then the resulting decision-making algorithm may produce biased results.<sup>78</sup> Barocas and Selbst offer the example of an employer that relies on data about online expressions of interest to target its recruitment efforts.<sup>79</sup> Because of differences in access to broadband in different communities, relying on such data may cause an employer to underestimate the level of interest and qualifications in underrepresented communities. A recruiting strategy based on such data is likely to produce biased outcomes.

Barocas and Selbst identify several other mechanisms by which data models may produce biased outcomes. The process of “feature selection”—choosing which attributes to include in the analysis—can have “serious implications for the treatment of protected classes.”<sup>80</sup> If the attributes that explain variation within a protected class are not incorporated, the model may be unable to distinguish among members of the group, leading it to rely on broad generalizations that disadvantage individual members of the group.<sup>81</sup> Data models may also discriminate when neutral factors act as “proxies” for sensitive characteristics like race or sex.<sup>82</sup> Those neutral factors may be highly correlated with membership in a protected class, and also correlate with outcomes of interest.<sup>83</sup> In such a situation, those neutral factors may produce results that systematically disadvantage protected groups, even though the model’s creators have no discriminatory intent, and the sensitive characteristics have been removed from the data.<sup>84</sup> Finally, Barocas and Selbst point out that

---

76. See Barocas & Selbst, *supra* note 5, at 684.

77. Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 57 (2013); see also Crawford, *supra* note 5.

78. See Barocas & Selbst, *supra* note 5, at 684-86.

79. *Id.* at 685.

80. *Id.* at 688.

81. See *id.* at 689-90.

82. See *id.* at 691-92.

83. See *id.* at 691.

84. See *id.*

employers may use data models to intentionally discriminate against certain groups. Because data mining can often infer protected class status from other neutral variables, employers could use data analytics as cover for intentional discrimination.<sup>85</sup>

In addition to the mechanisms that Barocas and Selbst identify, other characteristics of data models raise particular concerns when employers rely on them to make personnel decisions. Contrasting data mining techniques with traditional social science methodologies illuminates the problems. Social scientists articulate theories about the world, develop hypotheses based on those theories, and then subject those hypotheses to rigorous empirical testing, often by using data.<sup>86</sup> Their goal is to understand and explain patterns observed in the world.<sup>87</sup> An important part of designing an empirical test is determining what population the data should be drawn from and what variables should be included in the statistical model.<sup>88</sup> The theory motivating the study informs each of these decisions and each decision is consequential for the accuracy of the results.<sup>89</sup>

Suppose a researcher has a theory that past military service makes employees more successful in managerial positions. Testing this hypothesis will require examining how military service and on-the-job success are related using data about a representative group of workers. Looking only at those two variables might suggest that military service is negatively associated with future job performance. But a social scientist would also want to include other variables that could independently influence job performance. Unless the researcher controls for these factors, the study might reach an erroneous conclusion—a problem referred to as “omitted variable bias.”<sup>90</sup> If military recruits are significantly less educated than the rest of the population, looking only at the relationship between service and later job performance could be misleading. Including a variable for education in the model might show that

---

85. *See id.* at 692-93.

86. *See* Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 19-20 (2002).

87. *See id.* at 20-21, 60-61.

88. *See id.* at 54-55, 99-102.

89. *See id.*

90. *See id.* at 78.

military service is in fact associated with better job performance, after controlling for an individual's level of education.

The concern about omitted variable bias can apply to sensitive characteristics like race and sex in some circumstances. To extend the example above, suppose that for African Americans military service is highly positively correlated with subsequent work performance, while for white workers it has a somewhat negative effect. If the dataset includes far more observations about white workers, then a statistical model that omits race as a variable might predict that workers with past military service are less successful employees, even though the opposite is true for African Americans. If an employer relied on the model to disfavor workers with military experience, then the failure to include race as a control variable would ultimately disadvantage African Americans.

The solution is not to throw every possible variable into the statistical model.<sup>91</sup> Including too many variables might also bias results, especially if some variables are highly correlated. In such a situation, real effects are obscured, suggesting that no relationship exists among variables that are in fact related. Thus, for a social scientist trying to accurately describe relationships and effects in the real world, choices about which variables to include are crucial. Because the results of a statistical model are very sensitive to those choices, the norms of social science dictate that researchers be transparent about their choices and justify them by reference to the theory motivating the study. Those norms also encourage data sharing, to allow other researchers to replicate the study, to further test the results, and to criticize and revise the findings when necessary.

In contrast, data mining is inductive and atheoretical.<sup>92</sup> Data miners have no particular theory they are trying to test, nor are they necessarily interested in explaining observed relationships between different variables. Instead, data mining exploits enormous datasets with thousands of variables to uncover whatever statistical correlations might exist in the data. With no motivating theory to

---

91. *See id.* at 79-80.

92. *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, *BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK* 12-14 (2013) (explaining that big data shifts the focus from discovering causal relationships to uncovering patterns in the data); Custers, *supra* note 56, at 16.



justify the choices made, it is difficult to assess whether the data relied on is sufficiently representative, or whether the appropriate variables have been included to ensure the accuracy of the model. And in the absence of data sharing or transparency about the choices made in constructing the model, others cannot test the robustness and validity of the results.

Concerns that a data model may systematically disadvantage traditionally protected groups cannot be resolved simply by eliminating protected characteristics like race and sex from the data. As Barocas and Selbst explain, other types of information that closely correlate with those protected characteristics may serve as proxies, producing the same results without expressly relying on those categories.<sup>93</sup> At the same time, the possibility of omitted variable bias means that excluding race and gender variables will sometimes increase the risk of bias by failing to capture relevant differences between groups. The remedy is therefore not to exclude or include variables for sensitive characteristics in every case.

Because data mining is concerned only with identifying relationships, the model's creators often do not know whether correlations that are uncovered represent genuine relationships between factors in the real world or are artifacts of the data mining process. Social scientists expend a great deal of effort trying to determine whether an observed relationship between variables is causal. Because of the difficulty of establishing causality through statistics alone, a claim that two variables are related is subject to retesting and constantly open to challenge. By contrast, data mining models make predictions based on the strength of the statistical correlation alone.

In some contexts, we may not care much about the limitations of data mining. For example, if a computer algorithm can correctly flag which purchases made on my credit card are fraudulent and notify me, it does not matter whether I, or my bank, understand which variables triggered the alert or why. The difference between correlation and causation becomes important, however, if employers are basing their decisions on these statistical relationships. Suppose, for example, that data mining shows a strong statistical relationship between intelligence and "liking" curly fries on Facebook.<sup>94</sup> An

---

93. See, e.g., *supra* notes 82-83 and accompanying text.

94. See Kosinski et al., *supra* note 62, at 5804.

employer seeking highly intelligent employees might justify reliance on that correlation in selecting employees, even if it has a racially disproportionate effect, on the grounds that intelligence is a relevant job criterion.<sup>95</sup> If, however, the variables are merely correlated and not causally related, there is no necessary connection between them, and the correlation may not hold in the future. An employer relying on the statistical correlation may continue to make decisions disadvantaging minority applicants, even after the statistical relationship no longer holds true. Although it may seem clear that “liking” curly fries is not causally related to intelligence, in other cases it will not be intuitively obvious whether a given correlation is meaningful or spurious. But the same risk is present—that the algorithm is relying on a factor that has a discriminatory effect but is not actually connected to job performance.

Another novel challenge posed by data mining models is their lack of transparency. Many algorithms are built using machine learning techniques, which do not require the human programmer to specify in advance which factors the model should consider or what weight each should be given. Instead, the computer constructs a model by exploiting the relationships it uncovers between variables in the data. These relationships may be quite complex, such that in some cases the resulting model is completely opaque, even to its creators. When such a model is relied on to screen or rank applicants, it obscures the basis on which employers are making ultimate employment decisions. This lack of transparency makes it difficult to know if any observed bias is simply a byproduct of justifiable business considerations or the result of flaws in the model’s construction.

A related concern is that mechanisms to improve the accuracy of predictive models may not work in the context of employment. Big data enthusiasts often defend the use of algorithms on the ground that if the predictions are inaccurate, the machine will “learn” over time, such that any errors will be eliminated.<sup>96</sup> To return to the

---

95. The study by Kosinski and others also found that “liking” “I Love Being a Mom” is predictive of low intelligence. *Id.* The discriminatory impact on women that would result from relying on that apparent correlation is obvious.

96. See, e.g., MAYER-SCHÖNBERGER & CUKIER, *supra* note 92, at 12 (arguing that big data systems can “improve themselves over time” by continuing to look for signals and patterns as they receive new data).

example of credit card fraud detection, if the algorithm makes an error in classifying a charge on my credit card, I will discover the error sooner or later and report it. My feedback will be incorporated and the model will update and refine its decision process, becoming more accurate over time.

When applied to employment decisions, however, the process of error detection and learning is far less likely to occur. In the case of credit card fraud, consumers can observe and report the error. In the case of employment decisions, not all types of errors will be observable. Suppose an employer relies on an algorithm to sort applicants into “qualified” and “unqualified” pools. After hiring an applicant, the employer can observe the new employee’s work performance and will learn if the model made a mistake in classifying the applicant as qualified. However, if the algorithm mistakenly labeled an applicant as “unqualified,” the employer will not hire her and therefore will never observe her work performance. As a result, there will be no opportunity to learn of the error and update the model.

Once bias enters the system, feedback loops may reinforce that bias. Recall the example of Google’s algorithm which advertised criminal background checks more often when searches were conducted for African American-associated names than for Caucasian-associated names.<sup>97</sup> Those results likely reflect patterns in past search behavior, rather than any discriminatory bias on the part of the programmers who created the algorithm.<sup>98</sup> Nevertheless, the ads might nudge even the nonprejudiced employer, who otherwise would not treat applicants differently because of race, to scrutinize the criminal history of African American applicants more closely than white applicants. If, as a result of the nudge, the employer conducts criminal background checks more often for African American applicants than for white applicants, it will find more instances of criminal history in that population, further reinforcing a cycle of bias.

Feedback effects could also reinforce biased outcomes if disfavored groups are aware of the bias. If members of a particular group perceive that selection processes are systematically biased against

---

97. See Sweeney, *supra* note 17, at 46-47.

98. See *id.* at 52.

them and their chances of success are much less than for others, they may reduce their investment in developing their human capital.<sup>99</sup> This risk may be particularly significant if the patterns they observe suggest that the types of signals they have some control over—education, training, and the like—are not decisive and that other unknown or uncontrollable factors are shaping their employment opportunities.

Data mining models are thus far from neutral. Choices are made at every step of the process—selecting the target variable, choosing the training data, labeling cases, determining which variables to include or exclude—and each of these choices may introduce bias along the lines of race, sex, or other protected characteristics. Because of the atheoretical nature of data mining, once these biases are introduced, they may be difficult to detect and eliminate. Mere correlation may be mistaken for causation, and the true basis for employer decision-making is obscured. Moreover, these biases may persist or even worsen over time because of limited opportunities for error detection and the operation of feedback effects. For all of these reasons, identifying and addressing the potential harms that biased algorithms cause should be matters of policy concern.

### *C. Types of Harm*

Although many scholars have raised alarms that data analytics can produce biased outcomes, they have not articulated the precise nature of the harms that biased algorithms impose, or explained why they should be matters of policy concern. A common assumption among critics is that any type of bias in an algorithm is normatively troubling and requires policy or legal interventions. However, this assumption is unwarranted and overly broad. Virtually any decision-making process will produce disproportionate effects, and sometimes those effects will fall along protected class lines. What matters are the reasons unequal outcomes are occurring and whether those reasons are normatively acceptable.

---

99. See Samuel R. Bagenstos, *Subordination, Stigma, and "Disability,"* 86 VA. L. REV. 397, 464 & n.254 (2000) (citing economics literature that discrimination can be self-perpetuating if it discourages members of groups facing discrimination from investing in their human capital).

In this Section, I explain why certain types of bias in data models produce cognizable harms. Barocas and Selbst's taxonomy, discussed in the last Section, sought to explain the specific technical issues that can cause data models to discriminate.<sup>100</sup> My focus in this Section is different—namely, to identify the different types of harm that might result when employers rely on biased data models. Because the nature of the harm depends in part on the source of the bias, my typology of harms partially overlaps, but does not coincide, with their taxonomy. In what follows, I identify four distinct types of equality harms that may occur when employers rely on data analytics to distribute employment opportunities.

### *1. Intentional Discrimination*

One type of harm results when an employer uses data analytics to intentionally discriminate against a protected group.<sup>101</sup> In such a scenario, the employer relies on an algorithm to make hiring or promotion decisions because it *knows* the model produces a discriminatory result and *intends* that result to occur. The discriminatory decision simply masquerades behind the neutral façade of data analysis.<sup>102</sup> This type of discrimination is familiar as a form of intentional disparate treatment, only with the twist that the pretext—the “legitimate business reason” given for the decision—is the output of a computer model.

Although an employer might use data analytics as a screen for race or sex discrimination, an algorithm may be particularly effective in masking discrimination where the protected characteristic is not readily observable—for example, genetic traits and some kinds of disabilities. The law currently attempts to prevent these types of discrimination by restricting access to information about the protected characteristics. Thus, the Americans with Disabilities Act restricts an employer's ability to conduct medical exams or to

---

100. See Barocas & Selbst, *supra* note 5, at 677.

101. In Barocas and Selbst's taxonomy, this is referred to as “masking.” *Id.* at 692. Other scholars also catalogue the different ways that an algorithm can enable intentional discrimination. See, e.g., Dwork & Mulligan, *supra* note 5, at 36-38; Kroll et al., *supra* note 5 (manuscript at 32-34).

102. See Custers, *supra* note 56, at 9-10.

inquire about a disability prior to making a job offer,<sup>103</sup> and the Genetic Information Nondiscrimination Act forbids employers from seeking any kind of genetic information about applicants or employees.<sup>104</sup> An employer who believes that certain individuals are more costly to employ might use data profiles to identify and screen them out without ever explicitly asking for medical or genetic information. Several years ago, Target Stores used purchasing information to identify consumers who were in the early stages of pregnancy in order to send them coupons for baby products.<sup>105</sup> An employer with access to large amounts of behavioral data might similarly use that information to predict which applicants or employees pose future medical risks.<sup>106</sup>

When employers use data simply to mask intentional discrimination, the individual who loses out on an employment opportunity suffers the same type of harm as any other victim of intentional discrimination. The harm is direct and specific to the individual with the targeted characteristic.

## 2. Record Errors

A second type of harm arises when errors in an individual's record lead to the denial of an employment opportunity. For example, data collected from public sites might suggest that an individual has a criminal record or has defaulted on a loan, when in fact that is not true. The privacy literature, discussed in Part II.B, has focused on this type of harm. Inaccurate information does not inherently raise equality concerns, as errors may be randomly distributed, infecting the records of members of privileged groups as well as protected groups. However, evidence suggests that errors are more likely to

---

103. See Americans with Disabilities Act of 1990 (ADA) § 102, 42 U.S.C. § 12112(d)(2)(A) (2012) (“[A] covered entity shall not conduct a medical examination or make inquiries of a job applicant as to whether such applicant is an individual with a disability or as to the nature or severity of such disability.”).

104. See Genetic Information Nondiscrimination Act of 2008 (GINA) § 202, 42 U.S.C. § 2000ff-1(b) (“It shall be an unlawful employment practice for an employer to request, require, or purchase genetic information with respect to an employee or a family member of the employee.”).

105. See Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html> [<https://perma.cc/88XA-HHT7>].

106. See Zarya, *supra* note 10.

occur for members of subgroups that are farther from the mainstream. For example, individuals whose names have less common spellings—most likely ethnic names—have greater rates of error in records relating to their employability.<sup>107</sup> Similarly, people with two surnames—disproportionately Hispanics—or who have changed their names—disproportionately women—are more likely to have inaccuracies in their records.<sup>108</sup>

When an algorithm makes a prediction based on error-ridden data about an applicant, it may unfairly deprive that individual of an employment opportunity. The overall operation of the model may be unbiased in the sense that it accurately predicts outcomes for individuals about whom it has reliable data. If, however, errors are not randomly distributed, then the model's predictions may be more likely to produce erroneous predictions for some, and could result in outcomes systematically biased against members of certain groups.<sup>109</sup> In such a situation, it is theoretically possible to identify individual victims who can be made whole by granting access to the opportunities they would have had absent the errors in their records.<sup>110</sup> Of course, significant practical challenges may make it difficult to detect when errors are present in an individual's records and to prove that they caused the adverse outcome. Although proof may be difficult, the harm is easily conceptualized—identifiable individuals have lost out on specific employment opportunities.

### 3. *Statistical Bias*

A third type of harm may result from data models that are statistically biased, in the sense that they systematically disfavor a protected class because of the way the underlying model was created. Social scientists refer to statistical bias when problems

---

107. See AM. CIVIL LIBERTIES UNION, PROVE YOURSELF TO WORK: THE 10 BIG PROBLEMS WITH E-VERIFY (2013), [https://www.aclu.org/files/assets/everify\\_white\\_paper.pdf](https://www.aclu.org/files/assets/everify_white_paper.pdf) [<https://perma.cc/9K8A-N8L5>].

108. See EXEC. OFFICE OF THE PRESIDENT, *supra* note 22, at 52.

109. If record errors are pervasive for certain protected classes in the training data, they may also bias the model as a whole, such that even when applied to a population for whom accurate records are available, the outcomes will be biased. See Barocas & Selbst, *supra* note 5, at 684-85. This is a type of statistical bias, discussed in Part I.C.3.

110. Cf. Citron & Pasquale, *supra* note 20, at 4-5; Crawford & Schultz, *supra* note 20, at 101.

such as selection effects or omitted variables cause a model to be biased in the sense that it is systematically inaccurate in some way.<sup>111</sup> Similarly, data mining models built using biased, error-ridden, or unrepresentative data may be statistically biased.<sup>112</sup> Because of problems with the data or the model's construction, an algorithm may inaccurately capture relationships in the data, leading to imprecise or even erroneous predictions.

When statistical bias coincides with systematic disadvantage to protected classes, it causes discriminatory harm. The algorithm's creators may not have made the choices that produced the discriminatory effects with conscious intent to discriminate or even awareness of their biasing effects. Nevertheless, the resulting outcomes are not only biased in a statistical sense, but also in the colloquial sense of unfairly disadvantaging members of protected groups. The employer's practice has a discriminatory effect, and the statistical unreliability of the model undermines any justification for its use.

This type of bias, which results from the operation of a model, is structural in nature rather than individual. Correcting errors in the data about particular individuals will not solve the problem. Even if all the data used to predict future cases are entirely accurate, the algorithm produces results that are systematically biased against a protected group. The harm is also structural in the sense that it cannot be corrected for just one individual applicant or employee. The harmful effects on a protected group result from the operation of the model as a whole. This means that it may be difficult, if not impossible, to identify specific individual victims of discrimination.

Imagine a situation in which an employer relies on a biased algorithm to hire 100 employees from a pool of 1000 applicants. Suppose that 200 of the applicants (20 percent) are African American, but the employer only hires five. Of the 195 African American applicants who were not hired, it will be difficult to determine who *would* have been hired if the employer had not used the biased algorithm. Doing so requires making assumptions about what the model or the decision process would have looked like if constructed without the biasing choices. The difficulty is that there is not likely

---

111. See GARY KING, ROBERT O. KEOHANE & SIDNEY VERBA, *DESIGNING SOCIAL INQUIRY: SCIENTIFIC INFERENCE IN QUALITATIVE RESEARCH* 28 (1996).

112. See Barocas & Selbst, *supra* note 5, at 684-87.



to be a single unbiased alternative. Because model creation entails so many choices, multiple unbiased or less biased alternatives are possible, each of which might have selected a different set of individuals from the applicant pool for hire.

In the absence of a clear baseline against which to compare the outcomes, it is difficult to say that a particular individual in a protected class has been harmed while another has not. The harm to any given individual might be more accurately characterized as a reduction in their probability of selection rather than the loss of a job.<sup>113</sup> This uncertainty in identifying individual harms does not mitigate the fact that the operation of the model overall threatens a social harm if its effects are to entrench the disadvantage that subordinated groups experience.

#### 4. *Structural Disadvantage*

Even in the absence of statistical bias, an algorithm may produce disproportionate effects on a protected class. It may accurately capture the relationships between various attributes in the data in a way that produces outcomes that systematically disadvantage certain groups.<sup>114</sup> Note that with data mining models using large datasets, it may be practically difficult, if not impossible, to rule out the possibility that statistical bias has caused the discriminatory effects. At least as a theoretical matter, however, it is possible that a model is not biased in the statistical sense, but its operation systematically disadvantages members of a protected class. It might do so because the members of the protected class in fact differ in some systematic way relevant to characteristics that the model is trying to predict.<sup>115</sup>

In such a case, whether a rejected applicant has been harmed depends upon societal judgments about the fairness of the model. And whether a model should be considered fair depends on what attributes it leverages to make its predictions and on the normative

---

113. For a similar argument that affirmative action programs should be understood as altering the odds of success rather than actually depriving any particular individual of an opportunity, see Pauline T. Kim, Essay, *The Colorblind Lottery*, 72 *FORDHAM L. REV.* 9, 12, 30-35 (2003).

114. See Barocas & Selbst, *supra* note 5, at 691.

115. See *id.*

acceptability of relying on those factors. Certain attributes may be sufficiently related to job performance that the law should allow employers to rely on them regardless of their impact. For example, a company might reasonably screen applicants for legal positions to ensure that they are licensed to practice law, even if that selection criterion disadvantages certain groups. Whether employers should rely on other criteria, such as credit scores or criminal record history, is far more debatable, and resolving those questions turns on contested normative judgments.

The nature of data mining complicates our ability to make these types of judgments. Algorithms based on machine learning may be agnostic about what qualities make a good employee, and the resulting model may be opaque as to how it is sorting applicants or employees. Alternatively, the quality or characteristic the model seeks to maximize (the target variable) may be clearly specified, but the algorithm is so complex that it is not possible to explain which factors drive the model's predictions. Even when the factors are identifiable, a pure data mining model will not reveal whether the relationships uncovered are causal or merely coincidental.<sup>116</sup> Thus, in addition to familiar debates about whether certain selection criteria are closely enough related to the job, data analytics raise new questions about whether the law should permit employers to rely on unknown or unexplained correlations when they have the effect of disadvantaging certain groups.

Consider a simple example. Suppose a model analyzing tens of thousands of observations finds that residents of certain zip codes tend to perform more poorly at a particular job. Because residence is often associated with race, the model may effectively screen out minority applicants at higher rates. The data and methods used to build the model may be unimpeachable, such that there are no concerns about statistical bias. Or, put differently, the available evidence might suggest that the correlation is a genuine one. Nevertheless, as a normative matter, relying on this association may be unacceptable, not only because residence does not measure ability, but also because our country has a long history of housing segregation along racial lines.

---

116. See Custers, *supra* note 56, at 16.

A more difficult question is raised if the algorithmic bias results from a factor less clearly identified with past racial harms. Suppose, for example, that an algorithm uncovers a strong statistical correlation between job performance and a seemingly arbitrary factor like what kind of automobile someone drives, but the effect of relying on that factor is to reduce opportunities for members of a minority group. Some models may be so complex that it is impossible to specify which factors influence the results, or what precise weights different factors have in determining the model's predictions. Without knowing the precise mechanism producing the outcome, it is impossible to judge whether it is normatively acceptable to rely on the factors it leverages.

Thus, when an algorithm produces structural disadvantage that is not caused by statistical bias, the nature of the harm is more difficult to characterize. In such a case, the model's disparate outcomes may reflect genuine differences between groups that are relevant to job performance, or it may simply be capturing arbitrary and meaningless correlations. Whether it causes social harm depends on which differences the model leverages to make its predictions and on contested normative judgments about the acceptability of relying on those factors. To the extent that a harm occurs, however, it is a group-based harm. As with discriminatory statistical bias, the disadvantage is structural, and therefore identifying particular individual victims will be difficult.

#### *D. Classification Bias*

As discussed in Part I.C, algorithmic decision-making can produce various types of harms for individuals or protected groups deprived of employment opportunities. Apart from the first type—intentional discrimination—these harms do not easily fit traditional notions of discrimination as motivated by prejudice or animus. And yet, the growing use of big data and data analytics in the workplace risks creating or reinforcing patterns of disadvantage and subordination that will be very similar in effect to more familiar forms of discrimination from the past.

These risks raise a concern about what I call “classification bias”—namely, the use of classification schemes that have the effect of exacerbating inequality or disadvantage along lines of race, sex,

or other protected characteristics. I use the term classification bias to emphasize concerns about inequality and disadvantage, and at the same time to underscore that this type of bias results from mechanisms that are quite distinct from familiar forms of discrimination. More specifically, classification bias is *data-driven*, which means that the traditional legal tools for responding to discrimination are in many ways inadequate, as discussed in Parts II and III below.

The term “classification bias” resonates with the data science literature, which identifies “classification” as one of several basic data mining techniques;<sup>117</sup> however, I do not use the phrase in any technical sense. Other data mining techniques that are used to sort and score workers may also systematically disadvantage certain groups. Thus, classification bias applies whenever an algorithm—regardless of its logical structure—systemically biases applicants’ or employees’ access to opportunities.

In speaking of classification bias, I do not mean to invoke what is sometimes referred to as “anticlassification” theory.<sup>118</sup> Scholars have long debated what principles underlie antidiscrimination law. Some scholars have argued that the guiding principle should be one of formal equality—namely, that the law’s protections extend only as far as forbidding employers from making decisions based on an individual’s race, sex, or other protected characteristics.<sup>119</sup> This perspective, sometimes referred to as the “anticlassification principle,” identifies discriminatory harm primarily in the use of classifications—like race—to make decisions.<sup>120</sup> Anticlassification theory stands in contrast to antistatutory theory, which aims to promote equality by redressing structures and practices that disadvantage historically subordinated groups, regardless of whether the employer expressly or intentionally relied on race or other

---

117. See, e.g., Toon Calders & Bart Custers, *What Is Data Mining and How Does It Work?*, in *DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY*, *supra* note 56, at 27, 31-34.

118. See, e.g., Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antistatutory?*, 58 U. MIAMI L. REV. 9, 10 (2003) (describing the anticlassification principle as holding that the government may not classify people on the basis of a forbidden category such as race and explaining that it exists in tension with an antistatutory principle).

119. See, e.g., William Van Alstyne, *Rites of Passage: Race, the Supreme Court, and the Constitution*, 46 U. CHI. L. REV. 775, 797-98, 809-10 (1979).

120. See Balkin & Siegel, *supra* note 118, at 10.

categories in making its decisions.<sup>121</sup> Like antisubordination theory, the concept of classification bias proposed here looks at the consequences of employers' decisions. By asking whether neutral classification schemes work to systematically deprive already disadvantaged groups of opportunities, it shares the concerns of antisubordination theorists.

In Part III, I examine to what extent antidiscrimination law can respond to concerns about classification bias. But first, in Part II, I consider two other possible responses and explain why they are likely inadequate to meet the challenges posed by data-driven discrimination.

## II. ALTERNATIVE SYSTEMS OF REGULATION

This Part explores whether market forces or privacy law protections can be relied on to eliminate classification bias, and concludes that neither approach is likely to successfully meet concerns about inequality raised by workforce analytics.

### A. *The Market Response*

Proponents of market-based solutions might argue that the growing use of data mining models in employment raises no particular concerns because employers will rely on them only if they are effective. Collecting and analyzing data is expensive and employers will not do so, or pay a third party to do so, unless the benefits exceed the costs. The promised benefit of workforce analytics is that they will save employers time and money when making personnel decisions and will produce a better workforce.<sup>122</sup> Rational employers will not rely on these tools if they do not actually help them hire and retain good employees, and, therefore, market forces should eliminate models that are biased.

Michael Lewis's book, *Moneyball: The Art of Winning an Unfair Game*, has contributed to the idea that data analytics can accurately predict performance. *Moneyball* tells the story of Billy Beane, the

---

121. See, e.g., *id.* at 9; Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 157-58 (1976); Lawrence, *supra* note 37, at 319-20.

122. See Bersin, *supra* note 3.

Oakland Athletics manager who built a competitive baseball team with a limited payroll.<sup>123</sup> By substituting statistical analysis for hunches, intuition, and conventional wisdom, Beane was able to identify undervalued ballplayers and recruit them at a fraction of the cost of their true worth.<sup>124</sup> Since then, statistical analysis has become a standard tool that major league baseball teams use to identify talent.<sup>125</sup> The lesson seemed to be that statistics can not only help identify talent, but that they succeed in doing so because they are more “objective” and can overcome traditional prejudices.<sup>126</sup>

The success of statistics in baseball scouting does not translate easily to more ordinary jobs, however. As Nate Silver points out, baseball is unique in that it “offers perhaps the world’s richest data set.”<sup>127</sup> Not only are there *lots* of data about almost everything that happens in baseball games, but also the nature of the sport permits the collection of objective measures of individual performance under well-specified conditions—for example, batting statistics in a given ballpark against a particular pitcher.<sup>128</sup> Statistics revolutionized baseball to the extent that it did “because of the sport’s unique combination of rapidly developing technology, well-aligned incentives, tough competition, and rich data.”<sup>129</sup>

---

123. See generally MICHAEL LEWIS, *MONEYBALL: THE ART OF WINNING AN UNFAIR GAME* (2003).

124. See *id.* at 18, 37-42, 127-29.

125. See NATE SILVER, *THE SIGNAL AND THE NOISE: WHY SO MANY PREDICTIONS FAIL—BUT SOME DON’T* 86-88 (2012).

126. See *id.* at 91-92.

127. See *id.* at 80.

128. *Id.* (“[A]lthough baseball is a team sport, it proceeds in a highly orderly way: pitchers take their turn in the rotation, hitters take their turn in the batting order, and they are largely responsible for their own statistics.”). This type of data is harder to come by in other professional sports in which statistics have had less of an impact to date. See Leigh Steinberg, *Changing the Game: The Rise of Sports Analytics*, FORBES (Aug. 18, 2015, 3:08 PM), <http://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/> [<https://perma.cc/WXV6-EDL4>] (noting that although use of data analytics in all professional sports has increased, it is harder to adapt analytics to basketball than baseball); Reeves Wiedeman, *The Sabermetrics of Football*, NEW YORKER (Sept. 23, 2011), <http://www.newyorker.com/news/sporting-scene/the-sabermetrics-of-football> [<https://perma.cc/7TRY-F3FE>] (discussing why baseball is “more receptive to stats” than football). Even in baseball, statistics have not eliminated the role of scouts, and successful teams today use a combination of quantitative and qualitative information. See SILVER, *supra* note 125, at 91-92, 99-101.

129. SILVER, *supra* note 125, at 106.

In the more ordinary workplace, data models are more likely to exhibit bias,<sup>130</sup> and market competition will not reliably eliminate them. First, biased data models may be accurate *enough* to persist in a competitive market, even though they are biased against certain groups. Second, feedback effects may appear to confirm the accuracy of biased data models, entrenching their use. And finally, biased data models may be efficient precisely *because* they are discriminatory, and therefore pressures toward efficiency will not eliminate them.

The first reason that market pressures are unlikely to drive out classification bias is that a model may be sufficiently accurate to benefit employers who use them, even if, at the same time, they have a discriminatory effect. Consider an algorithm that selects candidates who are predicted to be more successful at a particular job. It may be highly effective in identifying strong candidates, even though it disproportionately excludes members of disadvantaged groups. So long as the algorithm is accurate enough to make the employer's process less costly, neither the employer nor the vendor will have sufficient incentive to identify and remove the bias.

This difficulty is compounded when considering singular, high-level positions for which there are few objective measures of performance. In baseball, the availability of highly detailed, objective, and publicly available data about performance means that a team will have numerous observations for comparing the performances of players in nearly identical circumstances.<sup>131</sup> In the case of other highly skilled workers, comparing performance is far more difficult. Finding an objective measure may not be possible, and even if one exists, comparisons will be difficult because a firm cannot observe the performance of the accepted and rejected candidates under identical circumstances. Without this information, it is difficult to assess the benefit or the cost of the choice actually made.

Imagine a company that relies on a data algorithm to choose among several applicants for a management position. The model might be biased in a way that discounts the leadership styles more typical of female candidates, such that it systematically assigns

---

130. *See supra* Part I.B.

131. Offensive ability in baseball is reliably captured by statistics, but defensive ability has proven somewhat more challenging to measure objectively. *See* LEWIS, *supra* note 123, at 136.

them lower scores, but nevertheless accurately identifies some candidates who are capable of performing the job. The employer may not recognize that the model is biased—particularly if its predictions match the decision maker’s prior implicit assumptions or expectations. In other words, the same cognitive biases that data purportedly help to avoid may cause the human decision makers not to notice when the model is biased.

If, relying on such a model, the employer selects a man for the job, and that man is ultimately successful in the position, the employer will have no reason to question the algorithm, even though an unbiased model might have prioritized others, including more female candidates. A female candidate might also have been successful in the job—maybe even *more* successful—but the employer will have no way of knowing that. So long as the algorithm is accurate *enough*, the employer would have no reason to distrust it.

Employers may persist in using biased algorithms to select for low skill positions as well. For these jobs, the basic skills may be widely available in the labor pool, and the relevant performance metrics may be easier to measure and compare across time. For example, an employer concerned with high turnover in low-skilled positions can easily measure the length of job tenure of different employees. The employer may utilize data mining tools in an effort to select employees who will stay longer at the job, and then compare the job tenure of employees hired before and after adopting the model. If the employer observes that employees hired using the model stay on the job longer, it may take that as confirmation of its accuracy. In fact, the model may not have identified the factors that actually increase job tenure. Some other factor, such as a decrease in alternative employment options, may have caused the observed increase in job tenure and would have similarly influenced those applicants not selected to stay on the job longer as well. Alternatively, an unbiased model might have similarly increased tenure without the discriminatory impact. Nevertheless, the employer’s observations would not lead it to question the model, and it would likely continue to use it, even though the effect is to disproportionately screen out minority applicants.

A second reason market forces may not reliably squeeze out classification bias is that feedback effects may cause biased models to become more accurate over time—the model in effect becoming a



self-fulfilling prophecy. Suppose, for example, that an employer uses a data model to select employees for an entry-level position for which many applicants meet the minimum qualifications. If the model is biased, such that it overselects individuals in a dominant group, then fewer minority group members will be hired, and the employer will have little opportunity to observe their performance in the position. At the same time, members of the minority group—particularly if similar processes restrict their access to other employment opportunities—may perceive a lower return to effort and therefore lose the incentive to invest in learning relevant skills.<sup>132</sup> A model which erroneously underpredicted minority performance may become more accurate over time. If similar biases operate across multiple domains, affecting access to other critical resources like housing and credit, then these feedback effects will multiply. Thus, when biased selection processes create feedback effects, market forces will tend to affirm rather than disconfirm their usefulness.

Finally, in those cases in which a data model is accurate *because* it is discriminatory, market forces will not eliminate classification bias. As discussed earlier, a model may incorporate biased judgments—for example, ratings by supervisors that are themselves biased—as a measure of job performance. If employers use such a model to predict future cases, and the performances of the selected employees are then evaluated using the same biased measure, the outcomes will simply confirm the “correctness” of the model. In other situations, a model might capture real market differences between employees, but those differences are themselves the product of discriminatory forces. One can imagine, for example, that women are less productive in nontraditional employment settings if they face resistance to their presence that is manifested in harassment and noncooperation from their coworkers. A model that predicts future performance based on the past would both reflect prior discrimination *and* be highly accurate. Once again, an employer focused on efficiency gains is unlikely to abandon the model.

Thus, market forces will not reliably eliminate classification bias. The market may squeeze out highly inaccurate models that fail to provide enough benefit to justify the cost to employers. In many

---

132. See Bagenstos, *supra* note 99, at 464.

cases, however, algorithms are likely to have some predictive value even if they are biased against certain protected groups. If they are accurate enough, employers will not have strong market incentives to abandon them or to incur the costs of searching for less biased alternatives.

### *B. Privacy Rights*

If market forces will not reliably eliminate biased algorithms, then what about regulation aimed at protecting informational privacy? Can restrictions on the collection, disclosure, and use of personal information address the risks of classification bias that data analytics pose? Privacy law scholars argue for more robust rules regulating information flows, suggesting that such rules would not only protect dignitary and autonomy interests, but also address the risk of discrimination as well.<sup>133</sup> Although information rules can certainly mitigate some of the threats to workplace equality, they cannot entirely meet the challenges posed by workplace analytics. A full exploration of the complex relationship between privacy and discrimination is beyond the scope of this Article. Instead, this Section briefly explains why even robust privacy protections are unlikely to fully resolve concerns about data-driven discrimination in the workplace.<sup>134</sup>

In some circumstances, privacy rights can prevent intentional discrimination from occurring. Thus, antidiscrimination statutes sometimes incorporate restrictions on employers' information gathering. For example, the Genetic Information Nondiscrimination Act prohibits employers from inquiring about or otherwise deliberately acquiring genetic information about applicants and employees.<sup>135</sup>

---

133. See, e.g., Richards & King, *supra* note 20, at 409-13.

134. In earlier work, I argued that protecting the privacy of sensitive information could prevent genetic discrimination from occurring. See Pauline T. Kim, *Genetic Discrimination, Genetic Privacy: Rethinking Employee Protections for a Brave New Workplace*, 96 NW. U. L. REV. 1497, 1501-02 (2002). That argument turned on the goal of preventing intentional discrimination and the fact that unexpressed genetic characteristics are not identifiable through casual observation. See *id.* at 1517, 1521. My observations about the connections between privacy and discrimination in that context do not necessarily apply in a data-rich environment where the discriminatory outcomes may not be intentional.

135. Genetic Information Nondiscrimination Act of 2008 (GINA) § 202, 42 U.S.C. § 2000ff-1(b) (2012).

The Americans with Disabilities Act similarly limits employers' access to medical information that might reveal the existence of a disability at certain stages of the employment process.<sup>136</sup> This strategy works when the protected characteristic is not readily observable.<sup>137</sup> If the employer does not know about a protected characteristic, such as a disability or a genetic predisposition to disease, it cannot discriminate on that basis. This strategy will obviously be less successful in preventing discrimination on the basis of highly salient characteristics like race and sex. Title VII of the Civil Rights Act does not contain a similar prohibition on acquiring information, although employer inquiries—for example, about an employee's plans to have children—may raise an inference that a later adverse action was taken on a prohibited basis. Restricting access to information can be effective in preventing intentional discrimination when the employer would not otherwise know about the protected characteristic and therefore would be unable to act on that basis.<sup>138</sup>

However, restricting access to sensitive information is not likely to be effective in preventing classification bias that results from data analytic models. If the data being mined is rich enough, other seemingly neutral factors may closely correlate with a protected characteristic, permitting a model to effectively sort along the lines of race or another protected characteristic.<sup>139</sup> Factors such as where someone went to school or where they currently live may be highly correlated with race. Behavioral data, such as an individual's Facebook "likes," can also predict sensitive characteristics like race and sex with a high degree of accuracy.<sup>140</sup> Because other information contained in large datasets can serve as a proxy for race, disability, or other protected statuses, simply eliminating data on those characteristics cannot prevent models that are biased along these dimensions. On the other hand, the problem of omitted variable bias means that prohibiting the collection or use of sensitive data may

---

136. Americans with Disabilities Act of 1990 (ADA) § 102, 42 U.S.C. § 12112(d)(2)(A).

137. See Kim, *supra* note 134, at 1517; see also CAL. GOV'T CODE § 12940(d) (West 2016).

138. Cf. Kim, *supra* note 134, at 1521.

139. See Custers, *supra* note 56, at 9-10.

140. See, e.g., Woodrow Hartzog & Evan Selinger, *Big Data in Small Hands*, 66 STAN. L. REV. ONLINE 81, 83 (2013); Kosinski et al., *supra* note 62, at 5804 fig.4.

sometimes increase the biased effects of a data model.<sup>141</sup> Thus, a simple prohibition on access to sensitive information will not prevent classification bias, and in some cases could make it worse.

Another approach to protecting privacy focuses on procedural protections. Fair information practices emphasize the right of individuals to know when and how personal data is collected, to ensure its accuracy, and to consent to its use.<sup>142</sup> However, these procedural rights have not significantly limited the types of data collected or how employers use that information. Applicants and employees often have little choice but to acquiesce to employer requests for information, and the law grants employers wide discretion in making employment decisions.<sup>143</sup> As a result, the emphasis on consent and data accuracy has had limited practical effect in restricting the information available to employers to make employment decisions.

Experience with the Fair Credit Reporting Act (FCRA), which embodies fair information practice principles, is illustrative.<sup>144</sup> The FCRA requires an employer to obtain an applicant's consent before it accesses a consumer report,<sup>145</sup> to provide notice of an adverse action based on a consumer report along with a copy of the report, and to provide information about the individual's rights to dispute the report's accuracy.<sup>146</sup> These requirements put few obstacles in the path of employers who wish to use consumer data to make personnel decisions. Job applicants have little choice but to consent to the use of credit reports if they wish to be considered for a job. If an

---

141. See *supra* Part I.B.

142. See, e.g., EXEC. OFFICE OF THE PRESIDENT, *supra* note 22, at 17.

143. See, e.g., Pauline T. Kim, *Privacy Rights, Public Policy, and the Employment Relationship*, 57 OHIO ST. L.J. 671, 717 (1996).

144. See Fair Credit Reporting Act (FCRA) § 602, 15 U.S.C. § 1681 (2012). For a more detailed discussion of the FCRA's limited ability to address concerns about algorithmic bias, see generally Pauline T. Kim & Erika Hanson, *People Analytics and the Regulation of Information Under the Fair Credit Reporting Act*, 61 ST. LOUIS U. L.J. (forthcoming 2017), <https://ssrn.com/abstract=2809910> [<https://perma.cc/N35G-P9FR>].

145. The FCRA defines a "consumer report" as

[A]ny written, oral, or other communication of any information by a consumer reporting agency bearing on a consumer's credit worthiness, credit standing, credit capacity, character, general reputation, personal characteristics, or mode of living which is used or expected to be used or collected in whole or in part for the purpose of serving as a factor in establishing the consumer's eligibility for ... employment purposes.

15 U.S.C. § 1681a(d)(1).

146. *Id.* § 1681b(b).

employer denies employment based on the report, the applicant's recourse is to try to correct any errors in that record.<sup>147</sup> The FCRA provides no remedy against an employer for failure to hire even when the employer relied on an inaccurate credit report. Relying on an accurate record to make decisions violates no legal prohibitions either, as long as all of the procedural steps have been followed. Thus, fair information practice principles are unlikely to significantly limit employer use of data models.

Scholars have widely criticized the reliance on notice and consent to protect privacy interests, especially in the era of big data.<sup>148</sup> Lengthy, jargon-filled disclosures encountered in nearly every internet transaction do not provide real notice,<sup>149</sup> and because the alternative to accepting those terms is to refuse the service or transaction, consumers have little real choice about how their personal information will be handled. The processing of big data exacerbates the problem of obtaining meaningful consent. Separate data streams can be combined, and, once aggregated, data may reveal far more about an individual's habits, tastes, and opinions than the individual data points alone would suggest.<sup>150</sup> As the example of Target Stores predicting which consumers were pregnant demonstrates,<sup>151</sup> the disclosure of relatively trivial bits of information may reveal far more sensitive information when data is aggregated and analyzed. Thus, consent obtained at the moment data is collected is not meaningful, given that it is impossible to know all subsequent uses of that information and its impact in advance.<sup>152</sup>

In response to the challenges posed by big data, privacy scholars have proposed forms of regulation that go beyond traditional fair information practice principles. As Neil Richards and Jonathan King point out, privacy rules are not just about secrecy or restricting

---

147. *See id.* § 1681i(f)(2)(B)(i).

148. *See, e.g.*, Citron & Pasquale, *supra* note 20, at 27-28; Crawford & Schultz, *supra* note 20, at 108; Richards & Hartzog, *supra* note 20 (manuscript at 17-21); Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1880-81 (2013).

149. *See* Richards & Hartzog, *supra* note 20 (manuscript at 18) (citing studies).

150. *See* Solove, *supra* note 148, at 1889-90.

151. *See supra* note 105 and accompanying text.

152. Solove, *supra* note 148, at 1889-90.

access to personal information.<sup>153</sup> Rather, privacy should be understood as “the rules that govern how information flows.”<sup>154</sup> For example, Kate Crawford and Jason Schultz advocate for a form of procedural data due process entitling individuals to know when predictive analytics are used and to challenge the fairness of the process.<sup>155</sup> Danielle Citron and Frank Pasquale similarly assert that data subjects should have the right to correct inaccurate data and that regulatory oversight should ensure the fairness of scoring systems.<sup>156</sup>

Requiring data transparency, auditing for accuracy, and substantively regulating downstream uses of data are important steps in ensuring the fair use of data; however, these types of interventions cannot fully address the risk of classification bias in employment. Inaccuracies in an individual’s record may unfairly deprive her of a particular opportunity, but accurate records do not guarantee unbiased outcomes. If an individual is excluded because of errors in her individual record, procedural rights can help correct the errors. However, fixing errors in an individual’s record will not prevent statistical bias or structural disadvantage—harms which result from the overall operation, rather than any individual application, of an algorithm. Because these harms operate by reducing opportunities for members of a group as a whole, merely correcting individual errors will not eliminate them. Thus, even robust privacy law regimes that focus on data accuracy are likely insufficient to address concerns about classification bias in employment.

### III. THE ANTIDISCRIMINATION RESPONSE

If neither the market nor privacy protections can reliably prevent classification bias, what about antidiscrimination law? In the employment context, Title VII of the Civil Rights Act of 1964 was the landmark piece of legislation establishing the antidiscrimination norm by forbidding discrimination on the basis of race, color, religion, sex, and national origin.<sup>157</sup> Later federal enactments extended

---

153. See Richards & King, *supra* note 20, at 411-12.

154. *Id.* at 411.

155. See Crawford & Schultz, *supra* note 20, at 126-27.

156. See Citron & Pasquale, *supra* note 20, at 20-22.

157. See Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a) (2012).

protections to older workers<sup>158</sup> and individuals with disabilities,<sup>159</sup> and prohibited discrimination based on genetic traits.<sup>160</sup> How do these laws apply to bias that is data-driven? Barocas and Selbst examined this question and concluded that “Title VII would appear to bless” the use of algorithms, even when they work to disadvantage protected groups.<sup>161</sup> In this Part, I reject that conclusion, arguing instead that employment discrimination law can provide a vehicle for addressing classification bias, so long as the doctrine accounts for its data-driven sources. The discussion below focuses on Title VII, because both statutory text and judicial interpretation of other employment discrimination laws often follows that of Title VII.<sup>162</sup>

In Section A, I review the conventional understanding of Title VII which divides prohibited discrimination into two categories—disparate treatment and disparate impact—and explain its limitations in addressing classification bias. Section B argues that a close reading of the statutory text supports a finding that Title VII directly prohibits classification bias. In Section C, I consider what an effective legal response to classification bias might look like, and how it should differ from conventional disparate impact theory in order to more closely meet the unique challenges that biased algorithms pose. The last two Sections of this Part, D and E, consider whether there are any legal or practical limits to relying on antidiscrimination law to address classification bias.

### *A. The Conventional Account of Title VII*

Judges, litigants, and scholars commonly recite that Title VII prohibits two types of discrimination: disparate treatment and

---

158. See Age Discrimination in Employment Act of 1967 §§ 2-12, 14-15, 17, 29 U.S.C. §§ 621-634 (2012).

159. See Americans with Disabilities Act of 1990 §§ 2-4, 101-102, 42 U.S.C. §§ 12101-12112.

160. See Genetic Information Nondiscrimination Act of 2008 §§ 201-212, 42 U.S.C. §§ 2000ff to 2000ff-11.

161. See Barocas & Selbst, *supra* note 5, at 672.

162. Of course, there are differences between Title VII and the other antidiscrimination statutes, which might affect the analysis, but a close examination of Title VII is a reasonable starting point. Further work should explore the extent to which the arguments advanced here do or do not apply to prohibitions on discrimination based on age, disability, or genetic traits.

disparate impact.<sup>163</sup> The standard account holds that disparate treatment cases involve intentional discrimination based on a protected characteristic, whereas disparate impact cases target employer practices that are facially neutral but have discriminatory effects.<sup>164</sup> As many scholars have argued, this neat division of actionable discrimination into two discrete types oversimplifies the reality of how bias can operate in the workplace.<sup>165</sup> It also arguably oversimplifies the relationship between these types of discrimination as a doctrinal matter.<sup>166</sup> And, as I argue in Section B of this Part, it may not be the best reading of the statutory text, or even an entirely accurate explanation of current doctrine.

Nevertheless, the conventional understanding is the place to begin. Read narrowly, existing Title VII doctrine does not appear to match the particular risks to workplace equality that classification bias poses.<sup>167</sup> Only one of the types of harm identified in Part I.C.—intentional discrimination—easily fits within the conventional framework. When an employer intends to discriminate but relies on an apparently neutral data model to justify its decisions, the traditional disparate treatment doctrine clearly applies.<sup>168</sup> The

---

163. See, e.g., *Furnco Constr. Corp. v. Waters*, 438 U.S. 567, 569 (1978); Charles A. Sullivan, *Disparate Impact: Looking Past the Desert Palace Mirage*, 47 WM. & MARY L. REV. 911, 914 (2005) (“Early in its history, the Supreme Court adopted two definitions of the term [‘discriminate’]: ... disparate impact ... [and] disparate treatment.”).

164. See *Ricci v. DeStefano*, 557 U.S. 557, 577-78 (2009); *Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 986-87 (1988).

165. See, e.g., Green, *supra* note 48, at 92; Krieger, *supra* note 39, at 1164-65; David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 899 (1993); Sturm, *supra* note 48, at 461.

166. See, e.g., Jed Rubenfeld, Essay, *Affirmative Action*, 107 YALE L.J. 427, 436-37 (1997); George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2313 (2006); Stacy E. Seicshnaydre, *Is the Road to Disparate Impact Paved with Good Intentions?: Stuck on State of Mind in Antidiscrimination Law*, 42 WAKE FOREST L. REV. 1141, 1142 (2007).

167. Barocas and Selbst similarly concluded that Title VII is “not well equipped” to address the various discriminatory features of data mining. See Barocas & Selbst, *supra* note 5, at 694.

168. Under the familiar *McDonnell Douglas* burden-shifting framework, the plaintiff has the initial burden of establishing a prima facie case of discrimination. See *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 801-03 (1973). The employer must then “articulate some legitimate, nondiscriminatory reason” for the adverse employment action. *Id.* Finally, the plaintiff has the opportunity to show that the employer’s proffered justification is pretext for discrimination. *Id.* at 804. If an employer were to point to the predictions of a data model to justify an adverse decision, the plaintiff could try to prove that the model is merely a pretext for intentional discrimination.



plaintiff may find it quite difficult as a practical matter to prove the employer's discriminatory intent in using a biased data model;<sup>169</sup> however, this scenario poses no conceptual difficulties under the disparate treatment theory.

As discussed in Part I.B, simply prohibiting use of protected characteristics will not prevent classification bias. Other nonsensitive variables can act as proxies, such that a model that does not explicitly consider race or sex may nevertheless have discriminatory effects along those lines. Moreover, because of the problem of omitted variable bias, forbidding the use of protected class variables could exacerbate discriminatory effects under certain circumstances. Thus, a blanket prohibition on the explicit use of race or other prohibited characteristics does not avoid, and may even worsen, the discriminatory impact of relying on a data model.<sup>170</sup>

The other types of harm resulting from classification bias—due to individual record errors, statistical bias, and structural disadvantage—can occur without any conscious intent or awareness on the part of the employer. Disparate impact doctrine would thus seem the natural place to look for a response. First articulated by the Supreme Court in *Griggs v. Duke Power Co.*, the disparate impact theory holds that Title VII forbids not only overt discrimination, but also “practices that are fair in form, but discriminatory in operation.”<sup>171</sup> The *Griggs* Court held that Duke Power could not require applicants to have a high school diploma or a passing score on a written test unless those requirements had “a demonstrable relationship to successful performance.”<sup>172</sup>

The disparate impact theory recognized in *Griggs* was rooted in Title VII's purpose—“to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees.”<sup>173</sup> Given that purpose, the Court held that Title VII required “the

---

169. See Barocas & Selbst, *supra* note 5, at 712-14.

170. In fact, mitigating the risk of biased outcomes arguably requires *preserving* data on race, sex, and other protected characteristics. See *infra* Part III.C.

171. 401 U.S. 424, 431 (1971). The *Griggs* Court explained that “artificial, arbitrary, and unnecessary barriers to employment” that “operate invidiously to discriminate on the basis of racial or other impermissible classification” are forbidden unless they “bear a demonstrable relationship to successful performance” of the job. *Id.*

172. *Id.*

173. *Id.* at 429-30.

removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification.”<sup>174</sup> The lack of discriminatory intent did not absolve the employer, for it “does not redeem employment procedures or testing mechanisms that operate as ‘built-in headwinds’ for minority groups and are unrelated” to the worker’s ability to do the job.<sup>175</sup>

As described in *Griggs*, the disparate impact theory would appear well-suited to address classification bias. Reliance on algorithms will typically be a facially neutral employment practice. Data models that do not explicitly categorize on the basis of race or other protected categories may nevertheless operate as “built-in headwinds” for disadvantaged groups. However, since the Court first articulated the concept of disparate impact in *Griggs*, a doctrinal superstructure has developed around the theory, which does not fit well when bias is data driven.<sup>176</sup>

As refined in subsequent cases and eventually codified by the Civil Rights Act of 1991, disparate impact liability attaches when a plaintiff has shown that an employment practice produces a disparate impact on the basis of a protected characteristic and the employer “fails to demonstrate that the challenged practice is job

---

174. *Id.* at 431.

175. *Id.* at 432.

176. Numerous scholars have noted the limitations of the doctrine and its failure to meet initial expectations of its transformative potential. Civil rights advocates initially heralded the *Griggs* decision as monumentally important in advancing the cause of workplace equality. See, e.g., Robert Belton, *Title VII at Forty: A Brief Look at the Birth, Death, and Resurrection of the Disparate Impact Theory of Discrimination*, 22 HOFSTRA LAB. & EMP. L.J. 431, 433 (2005) (“Aside from *Brown v. Board of Education*, the single most influential civil rights case during the past forty years that has profoundly shaped, and continues to shape, civil rights jurisprudence and the discourse on equality is *Griggs v. Duke Power Co.*”); Alfred W. Blumrosen, *The Legacy of Griggs: Social Progress and Subjective Judgments*, 63 CHI.-KENT L. REV. 1, 1-2 (1987) (“Few decisions in our time—perhaps only *Brown v. Board of Education*—have had such momentous social consequences [as *Griggs*].” (footnote omitted)). However, many others have viewed the doctrine more skeptically, arguing that it has been narrowly applied, is inherently limited, and lacks a clear theoretical basis. See, e.g., Rutherglen, *supra* note 166, at 2314; Selmi, *supra* note 28, at 706 (“[D]isparate impact claims are more difficult—not easier—to prove than claims of intentional discrimination.”); Sullivan, *supra* note 163, at 970, 975-76; Amy L. Wax, *Disparate Impact Realism*, 53 WM. & MARY L. REV. 621, 626 (2011); Steven L. Willborn, *The Disparate Impact Model of Discrimination: Theory and Limits*, 34 AM. U. L. REV. 799, 804 (1985); Nicole J. DeSario, Note, *Reconceptualizing Meritocracy: The Decline of Disparate Impact Discrimination Law*, 38 HARV. C.R.-C.L. L. REV. 479, 484, 507 (2003).

related for the position in question and consistent with business necessity.”<sup>177</sup> Even if the employer satisfies this burden, complainants might still prevail by demonstrating the existence of a less discriminatory alternative.<sup>178</sup> More specifically, a complaining party could “show that other tests or selection devices, without a similarly undesirable ... effect [on the protected class], would also serve the employer’s legitimate interest.”<sup>179</sup>

Michael Selmi argues that the disparate impact doctrine is not well suited to application outside the contexts in which the doctrine developed.<sup>180</sup> He points out that the early cases focused on seniority systems and written tests that employers used to perpetuate discrimination that had been lawful prior to the passage of Title VII.<sup>181</sup> Contemporaneous commentators understood the significance of the *Griggs* case as defining what was required to validate written employment tests.<sup>182</sup> The next disparate impact case decided by the Supreme Court, *Albemarle Paper Co. v. Moody*, also involved a challenge to preemployment tests, as well as an employer’s seniority system.<sup>183</sup> According to Selmi, application of disparate impact doctrine to these practices was relatively straightforward because they involved “specific practices that were easy to identify and for which there was no difficult causal question” and “[t]he employers’

---

177. See Civil Rights Act of 1991 § 105, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

178. See 42 U.S.C. § 2000e-2(K)(1)(A)(ii).

179. See *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975). The exact standard for establishing liability based on the existence of an alternative employment practice is uncertain because rather than defining the standard, Congress in the Civil Rights Act of 1991 simply referred to “the law as it existed on June 4, 1989, with respect to the concept of ‘alternative employment practice.’” See 42 U.S.C. § 2000e-2(k)(1)(c). In effect, Congress restored the law as it existed before the Supreme Court’s decision in *Wards Cove Packing Co. v. Atonio*, decided on June 5, 1989, 490 U.S. 642 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1074, *as recognized in* *Tex. Dep’t of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc.*, 135 S. Ct. 2507 (2015). In doing so, Congress repudiated the Court’s suggestion in *Wards Cove* that “any alternative practices ... must be equally effective as [the employer’s] chosen hiring procedures in achieving [its] legitimate employment goals,” including factors such as cost and other burdens on the employer. See *id.* at 661. However, because there was disagreement prior to *Wards Cove* about what exactly was required to show the existence of an alternative employment practice, the Civil Rights Act of 1991 did not resolve the issue.

180. See Selmi, *supra* note 28, at 705.

181. See *id.* at 708-16.

182. See *id.* at 723.

183. See 422 U.S. at 408-09.

rationales were likewise relatively easy to define.”<sup>184</sup> When applied in other contexts lacking these characteristics, however, the doctrine does not fit well, and liability is far more difficult to prove. As a result, very few disparate impact cases have been successful outside of the specific contexts in which the doctrine developed.<sup>185</sup>

Similarly, traditional disparate impact doctrine is a poor fit for addressing classification bias. Most data models have none of the characteristics that Selmi identifies as making disparate impact doctrine workable. Rather than providing specific selection criteria that are justified by clearly stated rationales, data models typically involve opaque decision processes, rest on unexplained correlations, and lack clearly articulated employer justifications.

The written employment tests targeted in early disparate impact litigation were grounded in psychological theories regarding aptitude and ability.<sup>186</sup> These tests focused on identifying and measuring skills or personal characteristics relevant to successful performance of a job, and their validity could be evaluated in light of standards set by an established scientific discipline.<sup>187</sup> In contrast, data mining is entirely atheoretical.<sup>188</sup> The models exploit whatever data are available, rather than selecting which factors should be included or controlled for based on theoretical expectations.<sup>189</sup> As a result, if existing disparate impact doctrine is applied

---

184. Selmi, *supra* note 28, at 716. Once adopted, disparate impact doctrine came to be seen as a generalized method of proving discrimination in situations far removed from seniority systems and written tests. *See, e.g.*, *N.Y.C. Transit Auth. v. Beazer*, 440 U.S. 568, 584-85 (1979) (applying disparate impact doctrine to claim that a transit authority’s regulation prohibiting the use of narcotics by employees violated Title VII); *Dothard v. Rawlinson*, 433 U.S. 321, 328-29 (1977) (applying disparate impact doctrine to claim that height and weight requirements for employment discriminated against women).

185. *See* Selmi, *supra* note 28, at 739-43 (describing how intentional discrimination cases may be easier to prove, with many cases asserting claims under both disparate impact and disparate treatment doctrine, and succeeding on the disparate treatment claim but not on the disparate impact claim); *id.* at 753 (“[O]utside of the testing cases, there has been no area where the disparate impact theory has proved transformative or even particularly successful.”).

186. *See, e.g.*, *Ablemarle Paper Co. v. Moody*, 422 U.S. 405, 410-13 (1975) (challenging employer use of Revised Beta Examination and Wonderlic Personnel Testing); *Griggs v. Duke Power Co.*, 401 U.S. 424, 428-29 (1971) (challenging employer use of Wonderlic Personnel Test and Bennett Mechanical Comprehension Test).

187. *See* Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.15 (2016).

188. *See supra* note 92 and accompanying text.

189. *See supra* Part I.B.

mechanically, it will fail to address the mechanics underlying classification bias.

A couple of examples are illustrative. Under disparate impact doctrine, if a plaintiff shows that an employer practice has a disproportionate impact on a protected group, the employer may defend by showing that the practice is “job related ... and consistent with business necessity.”<sup>190</sup> If an employer could meet this burden simply by showing that an algorithm rests on a statistical correlation with some aspect of job performance, then the test is entirely tautological, because, by definition, data mining is about uncovering statistical correlations. Any reasonably constructed model will satisfy the test, and the law would provide no effective check on data-driven forms of bias. Similarly, in disparate impact cases courts tend to defer to employer judgments about what abilities or skills are necessary for a job when evaluating employer justifications for a practice.<sup>191</sup> However, data mining models often rely on “discovered” relationships between variables rather than measuring previously identified job-related skills or attributes. When the employer has not considered and clearly articulated the reasons for relying on particular criteria, it is unclear why any deference is warranted.

The differences between employment testing and data mining also mean that defenses based on section 703(h) of Title VII do not apply. That section excuses employers from liability for relying on “any professionally developed ability test” so long as the test is “not designed, intended or used to discriminate” on a protected basis.<sup>192</sup> Algorithms used to sort or score workers are not “ability tests” because they do not actually test ability—rather, they identify behavioral markers that appear to correlate with on-the-job success. The legislative history of section 703(h) indicates that Congress added it to the statute to immunize the practice—common at the time—of relying on standardized tests to select applicants for hire or promotion.<sup>193</sup> Reflecting this understanding, the Equal Employment

---

190. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

191. See Selmi, *supra* note 28, at 753; Wax, *supra* note 176, at 633-34.

192. 42 U.S.C. § 2000e-2(h).

193. The version ultimately adopted made clear that reliance on these types of tests was not permitted if “designed, intended or used to discriminate.” *Id.* The opinion in *Griggs* focused primarily on this language, adopting the Equal Employment Opportunity Commission’s

Opportunity Commission (EEOC) Uniform Guidelines on Employee Selection Procedures, which interpret section 703(h), rely on and incorporate standards regarding test validation established by the American Psychological Association.<sup>194</sup> Because the EEOC wrote them to address an entirely different practice, those Guidelines are simply irrelevant when evaluating the use of atheoretical data mining models that result in classification bias.

To be clear, the *concept* of disparate impact—the idea that facially neutral employer practices can have discriminatory effects—applies to classification bias. The problem is that the ways the doctrine has been applied in the past are not well suited to address the data-driven nature of classification bias. Disparate impact theory *can* meet these specific challenges; however, doing so will require some adjustments in how it applies to workforce analytics. Section C below explains what types of adjustments are required, but first I consider whether Title VII can be read to address classification bias directly.

### *B. A Closer Reading*

The conventional reading of Title VII assumes that disparate treatment and disparate impact exhaust the possibilities for proving a violation under the statute. Scholars concerned about implicit biases or workplace structures that disadvantage women or racial minorities have either argued that the disparate treatment or disparate impact theory ought to apply,<sup>195</sup> or expressed concern that neither theory fits.<sup>196</sup> Similarly, Barocas and Selbst's conclusion that Title VII “would appear to bless” the use of data models even when they produce discriminatory results<sup>197</sup> rests on the assumption that the only available alternatives are existing disparate treatment and disparate impact doctrines.

---

interpretation that section 703(h) requires that any test be “job related” and not merely professionally prepared. *See* *Griggs v. Duke Power Co.*, 401 U.S. 424, 436 (1971).

194. *See* *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 430-31 (1975).

195. *See, e.g.*, Green, *supra* note 48, at 145; Krieger, *supra* note 39, at 1231.

196. *See, e.g.*, Samuel R. Bagenstos, *Bottlenecks and Antidiscrimination Theory*, 93 TEX. L. REV. 415, 434-35 (2014) (reviewing JOSEPH FISHKIN, *BOTTLENECKS: A NEW THEORY OF EQUAL OPPORTUNITY* (2014)); Sullivan, *supra* note 163, at 1000.

197. *See* Barocas & Selbst, *supra* note 5, at 672.

Perhaps, however, these two doctrines do not exhaust the options for demonstrating the discrimination forbidden by Title VII. The operative language of section 703 is divided into two parts:

- (a) It shall be an unlawful employment practice for an employer—
- (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or
  - (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.<sup>198</sup>

The conventional reading of section 703 is that (a)(1) is about disparate treatment—which turns on motive<sup>199</sup>—whereas (a)(2) is about disparate impact—which focuses on discriminatory effects. This reading reflects the doctrinal superstructure that has devel-

---

198. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(1)-(2) (2012). The Age Discrimination in Employment Act and the Genetic Information Nondiscrimination Act contain nearly identical prohibitions. *See* Age Discrimination in Employment Act of 1967 § 4, 29 U.S.C. § 623(a)(1)-(2) (2012); Genetic Information Nondiscrimination Act of 2008 § 202, 42 U.S.C. § 2000ff-1(a)(1)-(2). The Americans with Disabilities Act (ADA) similarly forbids “limiting, segregating, or classifying a job applicant or employee in a way that adversely affects the opportunities or status of such applicant or employee because of ... disability.” *See* American with Disabilities Act of 1990 § 102, 42 U.S.C. § 12112(b)(1). However, the operative provisions of the ADA differ from Title VII in other significant ways—for example, by making unlawful an employer's failure to reasonably accommodate otherwise qualified individuals with a disability, and its use of “qualification standards, employment tests or other selection criteria that screen out or tend to screen out” individuals with disabilities. *See* 42 U.S.C. § 12112(b)(5)-(6). These differences may mean that the ADA applies to biased data models in different ways than Title VII—a discussion that is beyond the scope of this Article.

199. Although the Supreme Court has at times suggested that disparate treatment cases require proof of discriminatory motive, *see, e.g.,* *Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324, 335 n.15 (1977) (stating that “[p]roof of discriminatory motive is critical” in disparate treatment cases), subsection 703(a)(1) does not refer to “intent” or “motive” at all. Rather, interpretation of that provision hinges entirely on the words “because of.” As Noah Zatz argues, however, “because of” could be interpreted to mean many things other than “motivated by.” *See* Noah D. Zatz, *The Many Meanings of “Because Of”: A Comment on Inclusive Communities Project*, 68 STAN. L. REV. ONLINE 68, 68-69 (2015).

oped around Title VII rather than a coherent underlying theory of discrimination. As numerous scholars have pointed out, the distinction between disparate treatment and disparate impact is far from clear, and the two theories overlap quite a bit both conceptually and as a matter of proof.<sup>200</sup> Nevertheless, the notion that disparate treatment and disparate impact capture the entire meaning of subsections 703(a)(1) and (a)(2), respectively, is often an unquestioned assumption.

However, the conventional reading does not inevitably flow from the statutory language. Focusing on the text suggests that Title VII also forbids what I have called classification bias—namely, the use of classification schemes that have the effect of exacerbating inequality or disadvantage along lines of race, sex, or other protected characteristics. The language of section 703(a)(2) specifically refers to employer practices that “classify” employees in ways that “deprive or tend to deprive” individuals of employment opportunities because of protected characteristics.<sup>201</sup> Obviously, Congress did not have in mind the problem of biased data mining models when it enacted the language of section 703(a)(2) in 1964. Nevertheless, the language sweeps broadly enough to reach unanticipated employer practices that exacerbate or entrench inequality on prohibited bases.

Differences in the texts of subsections 703(a)(1) and (a)(2) support the conclusion that section 703(a)(2) has broader reach than section 703(a)(1). Section 703(a)(2) restricts an employer’s ability to “limit, segregate, or classify” its employees or applicants.<sup>202</sup> In contrast to section 703(a)(1), which focuses on actions, such as hiring, firing, setting compensation, or terms and conditions that are taken with respect to a particular employee, section 703(a)(2) focuses on group-based actions—limiting, segregating, or classifying—all actions that necessarily are taken along some generalizable dimension. Importantly, the prohibited actions are not defined as limiting, segregating, or classifying *on the basis* of race or other protected characteristics. Instead, the emphasis of the language is on actions (such as classifying) that “deprive or tend to deprive” employees of opportunities on a protected basis.

---

200. See, e.g., Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1343-44 (2010); Rutherglen, *supra* note 166, at 2322-23, 2325, 2327, 2329-30.

201. 42 U.S.C. § 2000e-2(a)(2).

202. See *id.*



Many courts and commentators have simply assumed that section 703(a)(2) is synonymous with disparate impact doctrine. However, the text of (a)(2) makes no mention of “disparate impact,” “discriminatory effects,” “business necessity,” or “job relatedness.”<sup>203</sup> These concepts are codified in section 703(k), leaving open the possibility that section 703(a)(2) has meaning beyond or apart from established disparate impact doctrine.

When the Supreme Court first articulated the disparate impact theory in *Griggs*, it was only loosely connected to the language of section 703(a)(2).<sup>204</sup> In framing the question presented—whether the Duke Power Company’s high school diploma and testing requirements were lawful under Title VII—the Court dropped a footnote citing to the language of section 703(a)(2).<sup>205</sup> The Court made no further mention of that particular statutory provision in the opinion. Instead, the Court rested its analysis on Congress’s objective in enacting Title VII—namely, “to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees.”<sup>206</sup> The only part of the text of Title VII that the Court engaged with at length was section 703(h), which permits employers to rely on professionally developed ability tests so long as they are not “designed, intended or used to discriminate.”<sup>207</sup>

Subsequent cases cited primarily to *Griggs* as authority for the disparate impact doctrine,<sup>208</sup> although the Court eventually explained that *Griggs* was grounded in the text of section 703(a)(2).<sup>209</sup>

---

203. *See id.*

204. *See Griggs v. Duke Power Co.*, 401 U.S. 424, 426 (1971).

205. *Id.* at 426 n.1.

206. *Id.* at 429-30.

207. *See id.* at 433 (emphasis removed) (quoting Civil Rights Act of 1964, Pub. L. No. 88-352, § 703, 78 Stat. 241, 257 (1964) (codified as amended in scattered sections of 42 U.S.C.)). Duke Power Company argued that section 703(h) authorized its use of general intelligence tests as a screening device. *See id.* Relying on guidance the EEOC had issued and the legislative history of section 703(h), the Court concluded that the employer could not rely on the provision to defend its testing requirement when the test was not job related. *Id.* at 433-36.

208. The next three disparate impact cases in the Supreme Court did not cite to section 703(a)(2) at all in the majority opinions. *See generally* *N.Y.C. Transit Auth. v. Beazer*, 440 U.S. 568 (1979); *Dothard v. Rawlinson*, 433 U.S. 321 (1977); *Albemarle Paper Co. v. Moody*, 422 U.S. 405 (1975).

209. *See Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 985-86 (1988); *see also* *Smith v. City of Jackson*, 544 U.S. 228, 235 (2005) (explaining that although *Griggs* “relied primarily on the purposes of the Act,” the Court subsequently found that the disparate impact theory

Some commentators questioned whether Title VII authorized disparate impact claims at all,<sup>210</sup> but those concerns became moot when Congress enacted the Civil Rights Act of 1991.<sup>211</sup> Congress passed that legislation in response to several Supreme Court decisions in the late 1980s that were widely criticized as interpreting the protections of Title VII too narrowly.<sup>212</sup>

One of those cases was *Wards Cove Packing Co. v. Atonio*, a disparate impact case involving two companies that operated salmon canneries in remote areas of Alaska.<sup>213</sup> The plaintiffs alleged that the employers' hiring and promotion practices had produced a racially stratified workforce, in which skilled jobs (noncannery jobs) were held predominantly by white workers, while unskilled jobs (cannery jobs) were held predominantly by nonwhites.<sup>214</sup> The Court of Appeals found a prima facie case of disparate impact, but the Supreme Court reversed, holding that the appeals court had relied on the wrong statistics to conclude that a disparate impact existed.<sup>215</sup>

---

"represented the better reading of the statutory text as well"); *Connecticut v. Teal*, 457 U.S. 440, 445-47 (1982).

210. *See, e.g.*, *Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc.*, 135 S. Ct. 2507, 2526 (2015) (Thomas, J., dissenting) ("[T]he foundation on which the Court builds its latest disparate-impact regime—*Griggs v. Duke Power Co.*—is made of sand." (citation omitted)); Nelson Lund, *The Law of Affirmative Action in and After the Civil Rights Act of 1991: Congress Invites Judicial Reform*, 6 GEO. MASON L. REV. 87, 94 (1997) (arguing that there was no basis for the Supreme Court's recognition of the disparate impact theory in *Griggs*); *see also* Selmi, *supra* note 28, at 708-24 (detailing the origins of the disparate impact cause of action).

211. *See* Pub. L. No. 102-166, 105 Stat. 1071 (1991) (codified as amended in scattered sections of 42 U.S.C.).

212. *See* Sullivan, *supra* note 163, at 961 ("In reaction to *Wards Cove* and other decisions issued during the 1988 Term of the Supreme Court, Congress passed, and President Bush signed, the Civil Rights Act of 1991.").

213. *See* 490 U.S. 642 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071, *as recognized in* *Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc.*, 135 S. Ct. 2507 (2015).

214. *See id.* at 647-48.

215. *See id.* at 655. The Court held that the Ninth Circuit Court of Appeals had erred by comparing the percentage of nonwhite workers in the cannery and noncannery positions, and concluding that the stark racial disparity between the two groups established a prima facie case of disparate impact discrimination. *Id.* The relevant statistical comparison, the Court explained, is "between the racial composition of [the at-issue jobs] and the racial composition of the qualified ... population in the relevant labor market." *Id.* at 650 (quoting *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 308 (1977) (alterations in original)). Because the cannery work force did not reflect the population of qualified workers for the noncannery jobs, the statistical disparity in racial composition between the two groups did not establish a

In remanding, the Court addressed several additional issues—arguably all dicta—regarding disparate impact litigation. First, it stated that plaintiffs must identify the specific employment practice that created the alleged disparate impact as part of the prima facie case.<sup>216</sup> Second, it lowered the burden placed on the employer to justify an employment practice—asking whether it “serves, in a significant way, the legitimate employment goals of the employer”<sup>217</sup> rather than whether it is job related or required by business necessity, as it had in earlier cases.<sup>218</sup> Finally, the Court reallocated the burden of proving the lack of a business necessity to the plaintiffs,<sup>219</sup> which made it more difficult for plaintiffs to establish liability by showing that a less discriminatory alternative existed that the employer failed to adopt.<sup>220</sup>

When Congress passed the Civil Rights Act of 1991, it responded to the Court’s decision in *Wards Cove* by codifying the disparate impact doctrine and overturning or rejecting some of the Court’s guidance on disparate impact cases. It did so by placing the burden on the employer to demonstrate that a challenged practice is “job related for the position in question and consistent with business necessity,”<sup>221</sup> and by making clear that if “the elements of a respondent’s decision-making process are not capable of separation for analysis, the decision-making process may be analyzed as one

disparate impact. *See id.* at 651.

216. *Id.* at 656-58.

217. *Id.* at 659.

218. *See, e.g.,* Dothard v. Rawlinson, 433 U.S. 321, 331 n.14 (1977) (finding that “a discriminatory employment practice must be shown to be necessary to safe and efficient job performance”); Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971) (stating that in disparate impact cases, “[t]he touchstone is business necessity”).

219. *See Wards Cove*, 490 U.S. at 659. After a prima facie case of disparate impact is established, “the employer carries the burden of producing evidence of a business justification for his employment practice,” but the ultimate burden of persuasion remains with the plaintiff. *Id.*

220. *See id.* at 661. The Court wrote that

[A]ny alternative practices which respondents offer up ... must be equally effective as [the employer’s] chosen hiring procedures in achieving [its] legitimate employment goals. Moreover, “[f]actors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer’s legitimate business goals.”

*Id.* (quoting *Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 998 (1988) (fifth alteration in original)).

221. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

employment practice.”<sup>222</sup> With regard to establishing liability by showing the existence of an “alternative employment practice,” the Act simply stated that such a showing “shall be in accordance with the law as it existed on June 4, 1989”—the day before the Supreme Court issued the *Wards Cove* decision—without trying to articulate the correct standard.<sup>223</sup>

Congress made these changes by adding a new subsection (k), which defined disparate impact liability, to section 703 of Title VII, and retaining the language of section 703(a)(2) intact. After the amendments, the statute continued to prohibit in section 703(a)(2) limiting, segregating, or classifying employees in ways that “deprive or tend to deprive any individual of employment opportunities” because of race, color, religion, sex, or national origin, separately from the prohibition in section 703(k) of employment practices that have a disparate impact. Thus, the Civil Rights Act of 1991 left open the possibility that the judicially elaborated theory of disparate impact, as codified in section 703(k), does not exhaust the meaning of section 703(a)(2).

Interestingly, dictum in the Court’s *Wards Cove* decision is consistent with a reading that gives section 703(a)(2) meaning apart from traditional disparate impact doctrine. Because the canneries operated on a seasonal basis in a remote location, the employers provided housing and meals. Cannery and noncannery workers were assigned to separate dormitories and mess halls, which resulted in racially stratified living and eating quarters. In passing, the Supreme Court commented that the racially segregated facilities could give rise to a separate claim under section 703(a)(2), apart from the plaintiffs’ claim of disparate impact in hiring and promotion.<sup>224</sup> The Court’s language is admittedly ambiguous, but one way

---

222. *Id.* § 2000e-2(k)(1)(B)(i).

223. *Id.* § 2000e-2(k)(1)(C).

224. More specifically, the Court clarified the reach of its opinion in a footnote:

The Court of Appeals did not purport to hold that any specified employment practice produced its own disparate impact that was actionable under Title VII. This is not to say that a specific practice, such as nepotism, if it were proved to exist, could not itself be subject to challenge if it had a disparate impact on minorities. *Nor is it to say that segregated dormitories and eating facilities in the workplace may not be challenged under 42 U.S.C. § 2000e-2(a)(2) without showing a disparate impact on hiring or promotion.*

*Wards Cove*, 490 U.S. at 655 n.9 (emphasis added).

In other words, even if no actionable disparate impact had produced the employer’s racially

of reading it is that section 703(a)(2)'s meaning is not cabined by the disparate impact doctrine.

In any case, the fact that Congress left section 703(a)(2) intact when it responded to *Wards Cove* in the Civil Rights Act of 1991 supports the idea that (a)(2) continues to have independent force apart from the traditional disparate impact theory codified in subsection (k). Without the doctrinal elaboration of disparate impact theory, the text of (a)(2) supports a finding that Title VII prohibits classification bias.

### *C. Addressing Classification Bias*

As discussed in Section III.B, Title VII could be read to directly prohibit classification bias when algorithms operate to systematically disadvantage protected groups. Alternatively, disparate impact doctrine might be adjusted in ways that address those concerns. In either case, an effective legal response will require developing the doctrine to meet the particular challenges posed by data-driven discrimination. This Section sketches what a legal prohibition of classification bias looks like and how it should differ from traditional disparate impact doctrine.

As a preliminary note, this exploration focuses on employer liability, leaving aside the question whether vendors who create these models and sell or license them to employers should bear any legal responsibility. Although Title VII does apply to employment agencies,<sup>225</sup> it is highly uncertain whether that provision reaches vendors. I do not attempt to answer that question here, focusing instead on how Title VII might be applied to employers to address classification bias caused by workplace analytics. Regardless of whether vendors are directly liable, employers who face potential legal responsibility will have an incentive to pressure vendors to avoid biased outcomes.

---

stratified workforce, the plaintiffs might still be able to use section 703(a)(2) to challenge the employer's use of a classification (cannery versus noncannery workers) that adversely affected the employees' status. In that case, the harm suffered by the workers was the segregated living and dining quarters, and the violation occurred because the employer relied on a neutral classification that had the effect of depriving individual workers of opportunities or status because of their race. See 42 U.S.C. § 2000e-2(a)(2).

225. See 42 U.S.C. § 2000e-2(b).

Prohibiting classification bias requires examining the actual impact of the algorithms used to sort applicants and employees, and asking whether they deprive individuals of employment opportunities along lines of race, sex, or other protected characteristics.<sup>226</sup> Like traditional disparate impact doctrine, classification bias focuses on facially neutral employment practices that have disproportionately adverse effects on disadvantaged groups.<sup>227</sup> And like disparate impact doctrine, classification bias is not concerned with employer intent or motive.<sup>228</sup> If an employer relies on a data-driven classification scheme to sort applicants or employees, then it should be responsible for the impact that selection device has on the opportunities of workers in protected classes.

Given the differing reasons that data analytics may produce biased outcomes, an effective legal response must differ from traditional disparate impact doctrine in a number of ways. First, the law should not require employers to purge sensitive information, such as race and sex, from datasets; instead, preserving such data is important to avoid bias. Second, the method of identifying the relevant labor market for statistical comparison should look quite different. Third, an employer's defense of an algorithm with biased effects should depend, not on a claim of job-relatedness, but on the employer proving that the underlying model is statistically valid and substantively meaningful. Fourth, unlike under traditional disparate impact doctrine, employers should be able to rely on a "bottom-line" defense.

### *1. Data on Protected Class Characteristics*

Understanding the sources of classification bias suggests quite different rules regarding information about protected characteristics such as race and sex. A formalist reading of Title VII might appear

---

226. See *id.* § 2000e-2(a)(2).

227. See *Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 988 (1988); *Connecticut v. Teal*, 457 U.S. 440, 446 (1982); *Griggs v. Duke Power Co.*, 401 U.S. 424, 429-30 (1971).

228. See *Watson*, 487 U.S. at 988 ("This Court has repeatedly reaffirmed the principle that some facially neutral employment practices may violate Title VII even in the absence of a demonstrated discriminatory *intent*." (emphasis added)); *Griggs*, 401 U.S. at 430 ("Under [Title VII], practices, procedures, or tests neutral on their face, and *even neutral in terms of intent*, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices." (emphasis added)).

to prohibit any use of variables capturing sensitive characteristics in a data model.<sup>229</sup> Certainly, a simple model that relied on race or other protected characteristics as the basis for adverse decisions would run afoul of Title VII's prohibitions. However, when dealing with a complex statistical model involving multiple variables, the appropriate treatment of these sensitive variables is more complicated. If the goal is to reduce biased outcomes, then a simple prohibition on using data about race or sex could be either wholly ineffective or actually counterproductive due to the existence of class proxies and the risk of omitted variable bias.<sup>230</sup> Instead, avoiding classification bias may sometimes call for excluding sensitive demographic variables and at other times call for *including* them. Any response to biased data models must be sensitive to these nuances.

Regardless of whether a particular model should include variables for protected characteristics, preventing classification bias requires that, at the very least, model creators preserve these data when they are already present in the training data.<sup>231</sup> If developers purge demographic variables such as race and sex from the dataset, it becomes more difficult, if not impossible, to determine whether a model is systematically biased. Preserving these variables allows a model to be tested to determine its effect on the distribution of opportunities among different groups. Thus, unlike standard readings of Title VII which might suggest that data on sensitive characteristics should be disregarded or deleted,<sup>232</sup> a focus on classification bias argues for preserving this data and using it to assess the risks that a model produces biased outcomes.

## 2. *Relevant Labor Market Statistics*

The Supreme Court has stressed the importance of identifying the correct labor pool for comparison purposes when using statistical

---

229. See Barocas & Selbst, *supra* note 5, at 694-95.

230. See *supra* Part I.B.

231. See Dwork & Mulligan, *supra* note 5, at 37 (arguing that having data about legally protected characteristics is necessary to avoid unintended biased outcomes); cf. Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.15 (2016) (requiring employers to maintain records and disclose the impact of tests and other selection procedures on employment opportunities).

232. See Barocas & Selbst, *supra* note 5, at 694-95.

evidence to establish disparate impact.<sup>233</sup> According to the Court, the “proper comparison [is] between the racial composition of [the at-issue jobs] and the racial composition of the qualified ... population in the relevant labor market.”<sup>234</sup> This requirement has led to conflicts in particular cases over how to define the comparison pool—for example, what indicia should be used to identify “qualified” applicants and what geographic area constitutes the “relevant labor market.”<sup>235</sup> How a court resolves these questions can determine whether complainants are successful in establishing a prima facie case of discrimination.<sup>236</sup>

The search for the proper comparator group makes sense when trying to diagnose whether an independently developed selection device, such as a written ability test, will have a disproportionate impact when a particular employer administers it. With data mining, however, the employment practice at issue—the predictive model—is derived from preexisting data about large numbers of individuals who are taken to be representative of the target population. By constructing the model from the data, the data miners implicitly assume that the dataset used to train the model is complete enough and accurate enough to identify meaningful patterns among applicants or employees. If the operation of the model on the training data demonstrates an adverse effect on a protected class, that showing should be sufficient to establish a prima facie case. A court should not require a complainant to collect additional

---

233. See, e.g., *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 650-51 (1989), *superseded by statute*, Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071, *as recognized in* *Tex. Dep’t of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc.*, 135 S. Ct. 2507 (2015); *Watson*, 487 U.S. at 997; *N.Y.C. Transit Auth. v. Beazer*, 440 U.S. 568, 585-86 (1979).

234. *Wards Cove*, 490 U.S. at 650 (quoting *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 308 (1977) (alterations in original)).

235. See, e.g., *Dothard v. Rawlinson*, 433 U.S. 321, 330 (1977) (“The appellants argue that a showing of disproportionate impact on women based on generalized national statistics should not suffice to establish a prima facie case.... There is no requirement, however, that a statistical showing of disproportionate impact must always be based on analysis of the characteristics of actual applicants.”); *In re Emp’t Discrimination Litig. Against Ala.*, 198 F.3d 1305, 1312 (11th Cir. 1999) (“The focus during this first stage of the inquiry, and indeed during the whole of the disparate impact analysis, is on defining the qualified applicant pool.”).

236. See, e.g., *Peightal v. Metropolitan Dade County*, 26 F.3d 1545, 1554-55, 1557 (11th Cir. 1994); *Maddox v. Clayton*, 764 F.2d 1539, 1555 (11th Cir. 1985).



data about some relevant comparator pool to establish the adverse effects of the model.

On the other hand, even if a model does not exhibit discriminatory effects when run on the training data, that fact cannot be taken as conclusive evidence that outcomes will be unbiased when a particular employer applies the model in the real world. If the data relied on to build the model were not sufficiently representative or accurate, the model may be statistically biased in ways that systematically disadvantage certain groups when applied to actual applicants or employees. Thus, courts should also permit complainants to demonstrate that the operation of the model on real cases produces biased outcomes.

### 3. *Employer Justifications*

Under disparate impact doctrine, an employer may defend against a *prima facie* showing of disparate impact by demonstrating that the challenged practice is “job related ... and consistent with business necessity.”<sup>237</sup> The exact meaning of this phrase is ambiguous, and the standard has proven difficult to apply consistently in practice.<sup>238</sup> When applied to data analytics, however, it is difficult to make sense of the standard at all. When an algorithm relies on seemingly arbitrary characteristics or behaviors interacting in some complex way to predict job performance, the claim that it is “job related” often reduces to the fact that there is an observed statistical correlation. If a statistical correlation were sufficient to satisfy the defense of job-relatedness, the standard would be a tautology rather than a meaningful legal test. In order to protect against discriminatory harms, something more must be required to justify the use of an algorithm that produces biased outcomes.

As discussed in Part I.C, error-ridden, biased, or unrepresentative data, or improper specification of variables can introduce statistical bias, undermining the accuracy of a data model. When these statistical biases coincide with class membership, reliance on the model can harm members of protected groups. In order for claimants to diagnose whether statistical bias has infected an algorithm,

---

237. See Civil Rights Act of 1991 § 105, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012).

238. See Selmi, *supra* note 28, at 721-24; Wax, *supra* note 176, at 628, 631-36.

they would need access to the training data and the underlying model. The claimants would have to trace how the data miners collected the data, determine what populations were sampled, and audit the records for errors. Conducting these types of checks for a dataset created by aggregating multiple, unrelated data sources containing hundreds of thousands of bits of information would be a daunting task for even the best-resourced plaintiffs. In addition, the algorithm's creators are likely to claim that both the training data and the algorithm itself are proprietary information. Thus, if the law required complainants to prove the source of bias, they would face insurmountable obstacles.

Given these hurdles and the employer's superior access to information about the model's construction, employers should bear the burden of establishing the model's validity. The existence of a statistical correlation should not be sufficient. Instead, because the employer's justification for using an algorithm amounts to a claim that it actually predicts something relevant to the job, the employer should carry the burden of demonstrating that statistical bias does not plague the underlying model. In other words, the employer should have to defend the accuracy of the correlations it relies on by showing that no problems exist with the data or model construction that are biasing the results, and not simply by showing a statistical correlation in the existing data.

If an employer were able to satisfy this burden—if we could be certain that no statistical biases affected the model—should that be sufficient to justify reliance on an algorithm, even if it produces biased outcomes? In other words, should an employer be permitted to use a model that creates structural disadvantage if it is clear that it is not caused by statistical bias? Answering that question turns on the legitimacy of the employer's justification for using the model. And making that judgment requires knowing something about what the model is measuring and how it relates to the particular job. When applied to data analytics, however, two distinct problems arise. The first is the issue of interpretability. The second is the difficulty of distinguishing meaningful from spurious correlations.

The problem of interpretability arises because the atheoretical nature of data mining and the availability of unguided machine-learning techniques often make it difficult to know what factors are driving outcomes. An algorithm may be a "black box" that sorts

applicants or employees and predicts who is most promising, without specifying what characteristics or qualities it is looking for. It may, for example, be trained simply to look for applicants who resemble individuals hired in the past. Alternatively, the target variable might be clearly defined—as, for example, when an employer seeks employees who will maximize sales or have the longest job tenure—but it may not be possible to identify which particular attributes or variables are driving the algorithm or to determine how they are weighted.

Even when a model is interpretable, its *meaning* may not be clear. Two variables may be strongly correlated in the data, but the existence of a statistical relationship does not tell us if the variables are causally related, or are influenced by some common unobservable factor, or are completely unrelated. For example, one study found that employees who installed new web browsers on their computers rather than using preinstalled software stayed longer on the job.<sup>239</sup> But it is unclear why this correlation exists. It is possible, although unlikely, that not using the default browser makes an employee more dedicated. More likely, some unobserved attribute leads some individuals to choose a nonstandard browser, and also affects their longevity on the job. Or, it could be that the observed relationship between browser choice and productivity is entirely coincidental. Other correlations seem much more likely to be spurious—an artifact of the data mining process rather than a meaningful relationship—such as the apparent correlation between “liking” curly fries on Facebook and higher intelligence.<sup>240</sup>

Given the significant risks that biased algorithms will reproduce or entrench existing disadvantage, employers should bear the burden of justifying their use when they have disproportionate effects on protected groups. When a model is interpretable, debate may ensue over whether its use is justified, but it is at least possible to have a conversation about whether relying on the behaviors or attributes that drive the outcomes is normatively acceptable. When a model is not interpretable, however, it is not even possible to have

---

239. See Joe Pinsker, *People Who Use Firefox or Chrome Are Better Employees*, ATLANTIC (Mar. 16, 2015), <http://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/> [<https://perma.cc/4ZAA-LFLS>].

240. See Kosinski et al., *supra* note 62, at 5804.

the conversation. In such a case, the employer should not be able to justify its use merely because it captures a statistical relationship.

#### 4. *The Bottom-Line Defense*

Another way that Title VII doctrine should be adjusted is to allow employers a bottom-line defense when an algorithm is part of a larger selection process that is not biased overall. In 1982 the Supreme Court rejected the “bottom-line defense” in a disparate impact case, *Connecticut v. Teal*.<sup>241</sup> The plaintiffs in *Teal* alleged that their employer had violated Title VII by using a written exam that had a disparate impact on black employees as the first step in a promotion process.<sup>242</sup> Because black and white employees had significantly different passing rates, the proportion of black employees who continued to be eligible for promotion was much lower than that of white employees. When the employer later promoted some of these employees, it over selected black employees from among the eligible candidates. The end result was that 22.9 percent of the black employees who initially took the test were ultimately promoted, as compared with 13.5 percent of white employees.<sup>243</sup>

The employer argued that this “bottom-line” result, in which black employees were promoted at higher rates than white employees, should be a defense to the plaintiffs’ Title VII suit.<sup>244</sup> The Supreme Court, in a five-to-four decision, rejected the employer’s argument on the grounds that the goal of Title VII, as interpreted in *Griggs*, is “to achieve equality of employment *opportunities* and remove barriers” to equality.<sup>245</sup> In the Court’s view, the ultimate outcome of the promotion process was irrelevant because the plaintiffs’ claim was that they were denied “the *opportunity* to compete equally with white workers on the basis of job-related criteria.”<sup>246</sup> The Court also argued that the focus of the statute’s protection was the individual, not groups, and therefore, Title VII

---

241. 457 U.S. 440, 442, 452-56 (1982).

242. *Id.* at 443-44.

243. *Id.*

244. *See id.* at 452-53.

245. *See id.* at 448 (quoting *Griggs v. Duke Power Co.*, 401 U.S. 424, 429-30 (1971)).

246. *Id.* at 451.

required the employer to afford each applicant an equal opportunity to compete.<sup>247</sup>

Regardless of whether the Court's rejection of the bottom-line defense made sense given the facts in *Teal*, addressing classification bias calls for a different approach. It is possible that relying on certain elements or factors in a data model may tend to disadvantage a protected group, but those effects might disappear when they are part of a more complex model that allows for interactions among multiple factors. Thus, including race or sex as a variable might not cause an overall discriminatory effect at all. In some circumstances, including these variables might even make the model less likely to have a discriminatory effect—thereby contributing to a more equal bottom line. Similarly, the inclusion of some neutral variables may bias outcomes based on protected characteristics but will not always do so, depending on the overall structure of the model. Because isolating the effect of particular variables is difficult, treating the algorithm as an undifferentiated whole will often make sense. And if the algorithm's operation does not disproportionately exclude members of protected groups, then no discriminatory harm has occurred.

What if the operation of an algorithm produces biased outcomes, but the model's predictions are only one input in the employer's selection process, and, in the end, there is no disparate effect on a protected class? In that case, should the law still hold the employer responsible for relying on a biased data model as part of its process? In the context of workforce analytics, permitting a bottom-line defense makes sense. First, as discussed above, when dealing with algorithms plagued by statistical bias or reproducing structural disadvantage, the harm is systemic rather than individual.<sup>248</sup> Given that the central concern is with workplace systems that disadvantage certain groups, those concerns are alleviated when the operation of the system as a whole does not produce biased outcomes.

More practically, allowing employers a bottom-line defense is more likely to encourage equality-promoting uses of data. If employers are potentially liable for biased effects at each step of their

---

247. *Id.* at 453-56.

248. *See supra* Part I.C.

hiring or promotion process, they will have little incentive for self-examination or evaluation of the structural impact of their choices. Instead, they are likely either to ignore the risk that algorithms can cause bias or simply to cease using data analytics altogether. In contrast, a legal regime that permits a bottom-line defense will encourage employers to audit the impact of selection tools—including decision-making algorithms—on their workforce composition and to create processes that produce less biased results overall.

\* \* \*

Thus far, this Part has considered how the law should look different from existing doctrine in order to respond to the equality challenges posed by workforce analytics. As explained, the law will have to depart from traditional disparate impact doctrine in significant ways in order to respond effectively. It might do so by recognizing classification bias as a separate type of harm prohibited by Title VII, or, alternatively, by adjusting disparate impact doctrine to be more responsive to the particular risks posed by discriminatory algorithms. Whether framed in terms of a prohibition on classification bias or a revised disparate impact theory, the critical point is that data analytics differ significantly from the employer practices challenged in earlier cases, and thus require a legal response adapted to those particular risks.

#### *D. A Note on Ricci v. DeStefano*

The previous Section discussed how Title VII might be applied in ways better suited to meet the challenges to equality posed by workforce analytics. In this Section, I consider whether anything in existing Title VII doctrine would preclude such a development. More specifically, some commentators have interpreted the Supreme Court's decision in *Ricci v. DeStefano* as casting doubt on the viability of disparate impact theory—and by implication, any doctrine that looks at the disparate effects of employer practices.<sup>249</sup> These concerns raise the question: does the Court's holding in *Ricci* bar the development of Title VII doctrine in ways that can meet the

---

249. See, e.g., Primus, *supra* note 200, at 1344, 1363.

risks of classification bias? For reasons I explain below, I believe the answer is clearly “no.” And for the same reasons, Title VII—read as a whole—should pose no barrier to employers’ voluntary use of data analytics to try to diagnose and reduce structural forms of bias.

The dispute in *Ricci* arose when the City of New Haven, Connecticut, refused to certify the results of promotional exams.<sup>250</sup> After administering the written portion, the City realized that the exams would have a racially disparate impact if certified: virtually all of the promotions would go to white firefighters, even though a significant proportion of the candidate pool was black or Hispanic.<sup>251</sup> Concerned about a possible disparate impact lawsuit if it made the promotions, the City decided not to certify the results.<sup>252</sup> Some of the firefighters who believed that they would have been promoted sued the City.<sup>253</sup> These firefighters alleged that the City’s refusal to use the test results constituted a form of disparate treatment discrimination in violation of Title VII and the Equal Protection Clause because the City had considered the racial impact of the tests in making its decision.

In *Ricci*, the five-justice majority accepted the plaintiffs’ argument that the City’s decision to discard the test results violated Title VII’s disparate treatment prohibition with very little discussion.<sup>254</sup> The majority summarily rejected the district court’s reasoning that the City’s motivation of avoiding disparate impact liability did not constitute discriminatory intent. Writing for the majority, Justice Kennedy explained, “Our analysis begins with this premise: The City’s actions would violate the disparate-treatment prohibition of Title VII absent some valid defense.”<sup>255</sup> In the

---

250. See *Ricci v. DeStefano*, 557 U.S. 557, 562-63 (2009).

251. “Seventy-seven candidates completed the lieutenant examination—43 whites, 19 blacks, and 15 Hispanics. Of those, 34 candidates passed—25 whites, 6 blacks, and 3 Hispanics.” *Id.* at 566. The top ten candidates were eligible to fill eight vacant lieutenant positions. *Id.* All ten candidates were white. *Id.* “Forty-one candidates completed the captain examination—25 whites, 8 blacks, and 8 Hispanics. Of those, 22 candidates passed—16 whites, 3 blacks, and 3 Hispanics.” *Id.* The top nine candidates were eligible to fill seven vacant captain positions. *Id.* Seven of the candidates were white, and two were Hispanic. *Id.*

252. See *id.* at 562 (describing how the City threw out the examinations after some firefighters threatened to sue the City if it promoted firefighters on the basis of the tests).

253. *Id.* at 562-63.

254. See *id.* at 579-80.

255. *Id.* at 579.

majority's view, the fact that the City accounted for the racially disparate results made its decision a form of intentional discrimination, such that Title VII's disparate treatment and disparate impact prohibitions appeared to be in conflict.<sup>256</sup>

From this starting premise, the majority's analysis turned to whether the City had a lawful justification for taking the action it did. The Court rejected the City's argument that its good faith belief that using the exams would be a disparate impact violation justified discarding the test results.<sup>257</sup> The majority also rejected the plaintiffs' position that an employer may never take race-conscious actions even if the employer knows that it would otherwise violate disparate impact.<sup>258</sup> Instead, the majority concluded that the City must have "a strong basis in evidence to believe it will be subject to disparate-impact liability" to justify its actions.<sup>259</sup> Examining the record evidence, the majority concluded that New Haven lacked the requisite "strong basis in evidence," finding the exams "job related" and "consistent with business necessity."<sup>260</sup> The majority therefore held that discarding the test results violated Title VII.<sup>261</sup>

In the wake of *Ricci*, some commentators have suggested that disparate impact faces an existential threat.<sup>262</sup> If disparate treatment and disparate impact are in conflict, and if the Equal Protection Clause forbids disparate treatment, then is the disparate impact prohibition itself unconstitutional? Justice Scalia clearly

---

256. *See id.* at 579-80.

257. *Id.* at 581-82.

258. *See id.* at 580.

259. *Id.* at 585.

260. *See id.* at 587.

261. *Id.* at 592.

262. *See, e.g.,* Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities*, 101 CORNELL L. REV. 1115, 1126-27 (2016); Kenneth L. Marcus, *The War Between Disparate Impact and Equal Protection*, 2008-2009 CATO SUP. CT. REV. 53, 55; Eang L. Ngov, *When "The Evil Day" Comes, Will Title VII's Disparate Impact Provision Be Narrowly Tailored to Survive an Equal Protection Clause Challenge?*, 60 AM. U. L. REV. 535, 538-39 (2011); Primus, *supra* note 200, at 1343-44; Lawrence Rosenthal, *Saving Disparate Impact*, 34 CARDOZO L. REV. 2157, 2161-62 (2013); *see also* Richard A. Primus, *Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact After Ricci and Inclusive Communities*, in *TITLE VII OF THE CIVIL RIGHTS ACT AFTER 50 YEARS: PROCEEDINGS OF THE NEW YORK UNIVERSITY 67TH ANNUAL CONFERENCE ON LABOR* 295, 295-96 (Anne Marie Lofaso & Samuel Estreicher eds., 2015) (concluding in light of the Court's decision in *Inclusive Communities* that the statutory disparate impact standard will survive constitutional scrutiny given the current Court composition).



intended to signal a looming constitutional issue in his concurring opinion;<sup>263</sup> however, the rest of the Justices were content to argue the merits in *Ricci* on purely statutory grounds.<sup>264</sup> This approach is sensible because there is a vast difference between a constitutional prohibition on race-based state action and the conclusion that Congress cannot require employers to dismantle practices that operate as “built-in headwinds” for disadvantaged minority groups.<sup>265</sup> Despite the alarms, *Ricci* can easily be read as consistent with the continuing constitutionality of disparate impact liability under Title VII.<sup>266</sup> In *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc.*, the Supreme Court held that disparate impact claims are cognizable under the Fair Housing Act.<sup>267</sup> This decision suggests that the theory will likely remain viable even if subject to a direct constitutional challenge.<sup>268</sup>

Putting aside the constitutional question—as the Court did in *Ricci*—the question is whether prohibiting classification bias that results from data models would conflict with Title VII’s prohibition on intentional discrimination. The Justices in *Ricci* divided five to four over how to frame the question before the Court. While five Justices started from the premise that disparate treatment and disparate impact obligations were in conflict in the case,<sup>269</sup> the four dissenting Justices saw no conflict at all.<sup>270</sup> Justice Ginsburg, who authored the dissent, argued that the best reading of Title VII understands the disparate treatment and disparate impact theories as working in concert to achieve the statute’s purposes of “ending workplace discrimination and promoting genuinely equal opportu-

---

263. See *Ricci*, 557 U.S. at 594 (Scalia, J., concurring) (“[R]esolution of this dispute merely postpones the evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution’s guarantee of equal protection?”).

264. See *id.* at 576-78, 584 (majority opinion).

265. See *Griggs v. Duke Power Co.*, 401 U.S. 424, 432 (1971).

266. See *Primus*, *supra* note 200, at 1374-75 (arguing that disparate impact doctrine will survive constitutional challenge under two of three proposed readings of *Ricci*); *cf. In re Emp’t Discrimination Litig. Against Ala.*, 198 F.3d 1305, 1324 (11th Cir. 1999) (finding that Title VII’s disparate impact provisions are a valid exercise of Congress’s Fourteenth Amendment enforcement power).

267. 135 S. Ct. 2507, 2525 (2015).

268. See, e.g., Bagenstos, *supra* note 262, at 1127-28; *Primus*, *supra* note 262, at 295-96.

269. See *Ricci*, 557 U.S. at 580.

270. See *id.* at 624-25 (Ginsburg, J., dissenting).

nity.”<sup>271</sup> In the view of the dissenting Justices, the employer who rejects criteria that systematically disadvantage minorities “due to reasonable doubts about their reliability can hardly be held to have engaged in discrimination ‘because of’ race.”<sup>272</sup>

Thus, the Justices were closely divided on whether discarding New Haven’s promotional exams constituted disparate treatment. An even stronger case can be made that abandoning a data model that produces racially biased results is not a form of disparate treatment. Richard Primus argues that one plausible reading of *Ricci* is that the City’s actions constituted disparate treatment because they “adversely affected specific and visible innocent parties.”<sup>273</sup> Certainly Primus is right that protecting the expectations of the plaintiffs was a significant concern for the Justices in the majority. Justice Kennedy wrote that the City “create[d] legitimate expectations” in the firefighters who took the tests.<sup>274</sup> Some, he noted, “invested substantial time, money, and personal commitment in preparing.”<sup>275</sup> The problem arose because once the City established and announced the selection process, invalidating the test results upset legitimate expectations.<sup>276</sup> Justice Alito, in his concurrence, similarly emphasized the personal sacrifices that individual plaintiffs made to qualify for promotion—one firefighter hired someone to read and record the study materials because he was dyslexic, and another gave up a part-time job in order to study.<sup>277</sup>

---

271. *See id.* at 624.

272. *Id.* at 625.

273. Primus, *supra* note 200, at 1362. Some commentators have argued that the challengers were not in fact “victims” at all. *See, e.g., Ricci*, 557 U.S. at 608 (Ginsburg, J., dissenting) (“[The white firefighters] had no vested right to promotion.”); *see also* Mark S. Brodin, *Ricci v. DeStefano: The New Haven Firefighters Case & the Triumph of White Privilege*, 20 S. CAL. REV. L. & SOC. JUST. 161, 181, 202-12 (2011). Regardless, Primus is correct that the majority in *Ricci* viewed the challengers as victims because they relied on a process announced in advance. *See* Primus, *supra* note 200, at 1372-73.

274. *Ricci*, 557 U.S. at 583 (majority opinion).

275. *Id.* at 583-84.

276. *See id.* at 583-84, 593 (“The injury arises in part from the high, and justified, expectations of the candidates who had participated in the testing process on the terms the City had established for the promotional process.”). *Contra id.* at 630 (Ginsburg, J., dissenting) (“The legitimacy of an employee’s expectation depends on the legitimacy of the selection method.”).

277. *See id.* at 607 (Alito, J., concurring).

This reading of *Ricci*—that the disparate treatment violation occurred because the City’s action created “visible victims”<sup>278</sup>—is not only consistent with the language of the opinions, but it also best fits the statutory language. Title VII does not forbid any employer decision just because it is made with an awareness of race. Instead, it forbids “adverse employment actions” taken “because of an individual’s race.”<sup>279</sup> Unlike the situation in *Ricci*, prohibiting the use of a biased algorithm does not constitute a disparate treatment violation because there has been no adverse employment action. No employee has been deprived of a job to which he is entitled because no employee has any right or legitimate expectation that an employer will use any particular model. Because data mining models are atheoretical and typically based on past behavioral observations,<sup>280</sup> applicants are unlikely to know exactly which factors weigh into the model, and so they cannot argue that they relied on the process. The applicant who might have been selected if the employer had used a data mining model that it chose to discard is thus in an entirely different position from the white firefighters in *Ricci* who studied in reliance on the announced test. With no reliance interest and no entitlement that the employer use any particular model, employees who might have been hired if a biased model was used have no plausible claim that they have suffered discrimination.

Because disparate treatment violations occur only when employees’ legitimate entitlements are disrupted, nothing in *Ricci* precludes interpreting Title VII to prohibit classification bias, nor would the decision prohibit employer attempts to identify and avoid such bias. Barocas and Selbst thus overstate the matter when they suggest that any legislation directed at reducing biased models might “run afoul of *Ricci*.”<sup>281</sup> They argue that attempts to regulate data mining are problematic because diagnosing the impact of a model requires taking protected class characteristics into account.<sup>282</sup> As explained above, however, the problem in *Ricci* was not that the

---

278. See Primus, *supra* note 200, at 1345, 1369-75.

279. Civil Right Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(1) (2012).

280. See *supra* Part I.

281. See Barocas & Selbst, *supra* note 5, at 725.

282. See *id.* at 725-26.

City took action with an awareness of its racial impact, but that the action entailed adverse employment actions against identifiable persons. Merely being aware of the racial consequences of a selection process does not constitute disparate treatment. Similarly, an employer's efforts to understand the racial consequences of its processes in order to avoid bias does not violate Title VII.

Even the five Justices who disapproved of the City's actions in *Ricci* agreed on this point. As Justice Kennedy wrote, "Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."<sup>283</sup> And, of course, the only way to ensure that a test is fair regardless of race is to pay attention to race. The clear implication is that mere race-consciousness in developing a selection criterion is not a violation of Title VII. Rather, the Supreme Court has repeatedly emphasized that voluntary compliance by employers is "the preferred means of achieving the objectives of Title VII"<sup>284</sup> and "essential to the statutory scheme."<sup>285</sup> As the majority in *Ricci* recognized, unless employers can act to *avoid* practices that have a disparate impact, the voluntary compliance efforts that Title VII calls for would come "to a near standstill."<sup>286</sup>

Barocas and Selbst also erroneously suggest that *Ricci* poses an obstacle to crafting a remedy for biased classification schemes.<sup>287</sup> They argue that "[a]fter an employer begins to use the model to make hiring decisions, only a 'strong basis in evidence' that the employer will be successfully sued for disparate impact will permit corrective action."<sup>288</sup> However, nothing in *Ricci* prevents a court

---

283. *Ricci v. DeStefano*, 557 U.S. 557, 585 (2009); *id.* at 628-29 (Ginsburg, J., dissenting) ("This Court has repeatedly emphasized that the statute 'should not be read to thwart' efforts at voluntary compliance. Such compliance, we have explained, is 'the preferred means of achieving [Title VII's] objectives.'" (alteration in original) (internal citations omitted) (first quoting *Johnson v. Transp. Agency*, 480 U.S. 616, 630 (1987); and then quoting *Local No. 93, Int'l Ass'n of Firefighters v. City of Cleveland*, 478 U.S. 501, 515 (1986))).

284. *Id.* at 581 (majority opinion) (quoting *Local No. 93*, 478 U.S. at 515).

285. *Id.* at 583 ("The standard leaves ample room for employers' voluntary compliance efforts, which are essential to the statutory scheme and to Congress' efforts to eradicate workplace discrimination.").

286. *Id.* at 581.

287. See Barocas & Selbst, *supra* note 5, at 725-26.

288. *Id.* at 726.

from enjoining the use of a biased model, or an employer from voluntarily ceasing to use the discriminatory algorithm once that bias has been detected. The majority in *Ricci* objected to *undoing* the results of the test once the employer announced and administered it;<sup>289</sup> the Court did not require the City to continue using the test results to make future promotion decisions. To suggest otherwise would lead to the absurd result that an employer, who ordinarily has a great deal of discretion to change its selection processes or criteria, would suddenly be prohibited from changing a practice the moment it learned that it had a disparate effect on a protected group. Such an outcome would produce the exact *opposite* effect that Congress intended Title VII to have—namely, it would freeze into place employer practices that work to systematically disadvantage minority applicants and employees. The way to avoid such an absurd result is to recognize that acting prospectively to prevent classification bias is not a form of intentional discrimination.

A remedy limited to prospective relief is entirely consistent with *Ricci*. Because applicants and employees have no entitlement that an employer will continue to use any particular selection device,<sup>290</sup> the employer harms no one if it discards one practice in favor of a different one. Things would be more complicated if a remedy required the employer to fire current employees who were hired using a biased selection device, but that has not been the type of remedy required in successful disparate impact suits, nor should it be a remedy in cases of classification bias. For similar reasons, employers would not run afoul of Title VII by voluntarily avoiding models that produce biased results. An employer might not be permitted to fire an employee solely because she was selected using a biased data model. However, Title VII should not be read to prohibit the employer from ceasing to use that model once it discovers the bias.

### *E. The Limits of the Liability Model*

Prohibiting classification schemes that disadvantage protected classes is a promising avenue for addressing the equality concerns raised by workforce analytics. Such an approach is grounded in the

---

289. See *Ricci*, 557 U.S. at 585.

290. See *id.*

text of Title VII and consistent with the statute's purpose. Because the risks posed by workforce analytics stem from different sources than traditional forms of workplace testing,<sup>291</sup> it makes sense to tailor the doctrine to those particular risks rather than to mechanically apply the details of disparate impact doctrine that were developed in a different context. However, relying on the threat of legal liability to prevent classification bias has limitations as well.

In order to enforce its prohibitions on employment discrimination, Title VII relies on both individual and agency enforcement. After exhausting the administrative process, individual workers can file suit under Title VII and seek injunctive and monetary relief.<sup>292</sup> Under the current version of the law, a successful complainant is entitled to lost wages and other forms of equitable relief, compensatory damages, punitive damages (in cases in which the defendant acted with malice or reckless indifference), and attorneys' fees.<sup>293</sup> The law caps the total amount of compensatory and punitive damages based on the size of the employer.<sup>294</sup> This remedial structure is intended in part to incentivize aggrieved individuals to enforce the prohibition against employment discrimination.

In addition to individual suits, the EEOC also has enforcement powers.<sup>295</sup> The EEOC has authority to receive, investigate, and conciliate charges of discrimination under Title VII and other antidiscrimination statutes. In cases in which the EEOC has found cause to believe discrimination occurred but was unable to resolve the dispute through informal conciliation, the EEOC may choose to file suit on behalf of a complaining party.<sup>296</sup>

For several reasons, this scheme may be less effective at enforcing a prohibition on classification bias, as compared with other types of discrimination. First, as previously discussed, the harms that classification bias causes are structural rather than individual in nature.<sup>297</sup> Because the harms are more diffuse, individuals will find it extremely difficult to detect when a biased algorithm has

---

291. *See supra* Part I.B.

292. *See* Civil Rights Act of 1964 § 706, 42 U.S.C. § 2000e-5(f) (2012).

293. *See* Civil Rights Act of 1991 § 102, 42 U.S.C. § 1981a(a)-(b); 42 U.S.C. § 2000e-5(g).

294. *See* 42 U.S.C. § 1981a(b)(3).

295. *See id.* § 2000e-5.

296. *See id.* § 2000e-5(f).

297. *See supra* Part I.C.

produced an adverse outcome and to understand what caused the model to be biased. Even if these obstacles are overcome, the appropriate remedy would be structural in nature—namely, an injunction to revise or eliminate use of a biased model.<sup>298</sup> The reduced chance of receiving damages makes it less likely that individual employees will step forward to challenge instances of classification bias.

Individual complainants may not be reliable enforcers of a prohibition on classification bias for another reason. Detecting and pursuing claims of classification bias will be highly resource- and time-intensive. Even with a favorable legal regime, plaintiffs will need experts to determine whether data models are producing biased outcomes. Most individual plaintiffs will simply be financially unable to pursue such a case, particularly when the likelihood of a large damage award is slim.

The EEOC might step into the breach, as it often does, by litigating cases that have the potential for significant public impact, but that private litigants are unlikely to pursue.<sup>299</sup> Even if the EEOC makes these cases a priority, however, its limited resources will significantly constrain its efforts. Currently, the EEOC receives nearly 100,000 new charges annually and it faces a persistent backlog of charges.<sup>300</sup> The EEOC's current strategic priorities include cases involving systemic discrimination.<sup>301</sup> That focus would seem to encompass the structural harms threatened by employer reliance on biased data models. If the EEOC decides to prioritize cases involving workforce analytics, it would need to develop methods for detecting when data algorithms are producing discriminatory outcomes. Doing so would require a level of technical expertise and fiscal resources even beyond what is currently needed to tackle large scale systemic cases.<sup>302</sup>

Lowering the standards for establishing liability or increasing the available remedies could resolve the problem of insufficient incentives for private litigants to file suit. If the law swings too sharply

---

298. See *supra* Parts I.C, III.D.

299. See Pauline T. Kim, *Addressing Systemic Discrimination: Public Enforcement and the Role of the EEOC*, 95 B.U. L. REV. 1133, 1141-46 (2015).

300. See *id.* at 1144.

301. See *id.* at 1141-42.

302. Cf. *id.* at 1145-46.

in that direction, however, it may deter employers from attempting to understand whether their data tools have any disparate effects, and they may prefer instead to remain ignorant of any biases those tools may be causing. Alternatively, employers may cease using data models altogether, even though data analytics might help to diagnose and correct existing cognitive or structural biases. Thus, the goal of the law should not be to eliminate the use of all data analytics. Instead, the optimal legal regime would deter the use of biased data models while permitting or encouraging equality-promoting uses of data. The difficulty of balancing these two goals under Title VII suggests that policymakers may need to look beyond a backward-looking, liability-based regime and to consider other regulatory responses.

Fully exploring alternative regimes goes beyond the scope of this Article, but a few examples are illustrative. Technological innovations may make it possible to limit in advance whether a computer will produce an algorithm with a disparate effect on a protected class.<sup>303</sup> Another possibility would be to develop an ex ante regulatory regime to govern algorithms like the one currently used for premarket approval of drugs.<sup>304</sup> An appropriately structured approval process could ensure that data mining models are not statistically biased and that the social costs of using them do not exceed the benefits. Alternatively, a regulatory body might work to develop standards relating to data collection, integrity and preservation, and model validity, such that models that complied with these standards would have a presumption of legality.

None of these alternatives is simple or guaranteed to work, and all are likely to generate resistance. Implementing any of these solutions would require resolving difficult questions about what kinds of bias are unfair and how much should be tolerated. But

---

303. See, e.g., Sara Hajian & Josep Domingo-Ferrer, *Direct and Indirect Discrimination Prevention Methods*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 241, 247-51; Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Techniques for Discrimination-Free Predictive Models*, in DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 56, at 223, 229-35; Kroll et al., *supra* note 5 (manuscript at 35-45).

304. Cf. Andrew Tutt, *An FDA for Algorithms*, 68 ADMIN. L. REV. (forthcoming 2017) (manuscript at 20-25), <https://ssrn.com/abstract=2747994> [<https://perma.cc/2WCH-WMB9>] (arguing for a federal regulatory agency to ensure the safety and efficacy of algorithms before they are introduced in the market).



these types of efforts might offer some kind of safe harbor to employers who, acting in good faith, attempt to leverage data to remove bias from their personnel practices.

#### CONCLUSION

The data revolution is here to stay. Advances in computing power and the availability of massive amounts of data make it inevitable that employers will harness these tools to manage their workforces. Depending on how employers deploy these tools, data may enhance workplace fairness or exacerbate inequality. When these tools are used—not as guides or aids, but as gatekeepers to critical employment opportunities—they risk reinforcing existing patterns of disadvantage. Because of the nature of data mining techniques, employer reliance on these tools poses novel challenges to workplace equality and thus traditional doctrine will not suffice to address them.

Thinking in terms of classification bias offers a lens through which to better understand these challenges and to consider how to develop an appropriate legal response. Although the term may sound novel, a legal prohibition of classification bias is grounded in the text of Title VII and fully consistent with its purposes. Whether recognized as a distinct type of discrimination under Title VII or a species of disparate impact theory, classification bias offers a way for rethinking how antidiscrimination law should be tailored to respond to the unique challenges raised by data-driven forms of discrimination. Doing so is essential for Title VII's vision of workplace equality to continue to advance in the face of evolving threats.