REVISION: June 14, 2016

Running head: RACE, RISK & RECIDIVISM

Risk, Race, & Recidivism:

Predictive Bias and Disparate Impact

Jennifer Skeem

University of California, Berkeley

jenskeem@berkeley.edu

and Christopher T. Lowenkamp

Administrative Office, U.S. Courts

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.

Electronic copy available at: https://ssrn.com/abstract=2687339
Electronic copy available at: http://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339
Electronic copy available at: http://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

> "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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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.

Electronic copy available at: https://ssrn.com/abstract=2687339

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).[i]

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)

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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 <u>not</u> 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:

Electronic copy available at: https://ssrn.com/abstract=2687339

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-

Electronic copy available at: https://ssrn.com/abstract=2687339

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.[ii] 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

Electronic copy available at: https://ssrn.com/abstract=2687339

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 α=.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 (α=.66; Spearman-Brown Estimated α |10 items=.76)
- "Employment and education" includes highest grade completed, unstable recent work history, and currently unemployed (α=.47; Spearman-Brown Estimated α |10 items=.75)
- "Social networks" includes family problems, unmarried, and lack of social support (α=.47; Spearman-Brown Estimated α |10 items=.67)

Electronic copy available at: https://ssrn.com/abstract=2687339

- "Substance abuse" includes recent alcohol problems and recent drug problems (α=.38 Spearman-Brown Estimated α |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

Electronic copy available at: https://ssrn.com/abstract=2687339

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,

Electronic copy available at: https://ssrn.com/abstract=2687339

$\chi^2(1) = 94.46$; $p < 0.001$, $\phi = -0.05$). 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]

Electronic copy available at: https://ssrn.com/abstract=2687339

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) focusers 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.[iii]

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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 [χ^2 (1) = 10.64, *ns; Pseudo-R^2* Δ=0.00] or violent arrest, [χ^2 (1) = 0.28, *ns*; *Pseudo-R^2* Δ=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 [χ^2 (1) = 9.1, *ns; Pseudo-R^2* Δ=0.00], but adds modest incremental utility in predicting violent arrest, [χ^2 (1) = 16.93, p <.001; *Pseudo-R^2* Δ=0.00]. The odds ratios for race in Model Three are small and not statistically significant at our

Electronic copy available at: https://ssrn.com/abstract=2687339

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,[iv] 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*].

Electronic copy available at: https://ssrn.com/abstract=2687339

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 [χ^2 (1) = 29.42, $p < .001$], but not violent arrest [χ^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^2 Δ=0.00) and the interaction term was small (OR =1.04, $p < .001$, 99.9%

Electronic copy available at: https://ssrn.com/abstract=2687339

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 χ^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 [χ^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

Electronic copy available at: https://ssrn.com/abstract=2687339

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 χ^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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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.

Electronic copy available at: https://ssrn.com/abstract=2687339

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 (ϕ =-.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 bootstapping. 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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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'

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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).

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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.[v]

Conclusion

In light of our results, it seems that concerns expressed about risk assessment are exaggerated. To be clear, we are <u>not</u> 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

Electronic copy available at: https://ssrn.com/abstract=2687339

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.

Electronic copy available at: https://ssrn.com/abstract=2687339

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^		
Drug	46	47
Firearms	16	16
White Collar	17	18
Other	8	9
Violence	5	5
Property	5	5

^ Categories with less than 5% combined as other (i.e., sex offense, public order)

PCRA=Post Conviction Risk Assessment

Electronic copy available at: https://ssrn.com/abstract=2687339

Table 2. Predictive Utility of PCRA by Race

	Any Arrest			Violent Arrest		
Feature	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

Electronic copy available at: https://ssrn.com/abstract=2687339

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 X^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 X^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

Electronic copy available at: https://ssrn.com/abstract=2687339

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)

Electronic copy available at: https://ssrn.com/abstract=2687339

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^2 Change	X^2	OR, Interaction (Model 4)	99.9% CI	R^2 Change	X^2	OR, Race (Model 3)	99.9% CI
Any Arrest								
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
Violent Arrest								
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
*$p < .001$

Electronic copy available at: https://ssrn.com/abstract=2687339

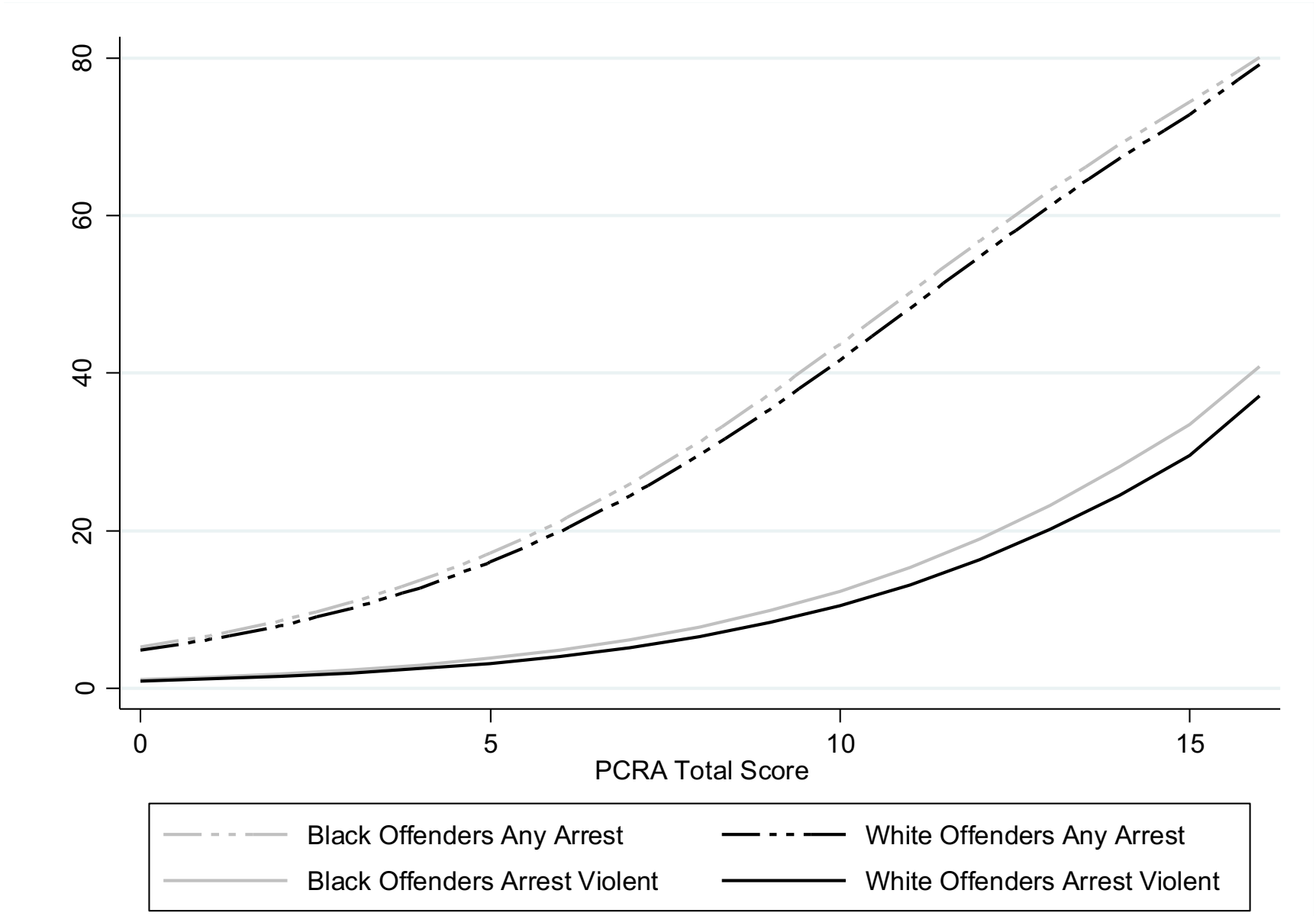
Table 6. PCRA Mean Score Differences by Race

	Black (N=16,537)		White (N=16,537)				Cohen's d		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Difference	% Attributable To	Estimate	Lower	Upper
PCRA Total	7.37	3.25	6.23	3.38	1.14		0.34	0.31	0.36
Domains									
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
Items									
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

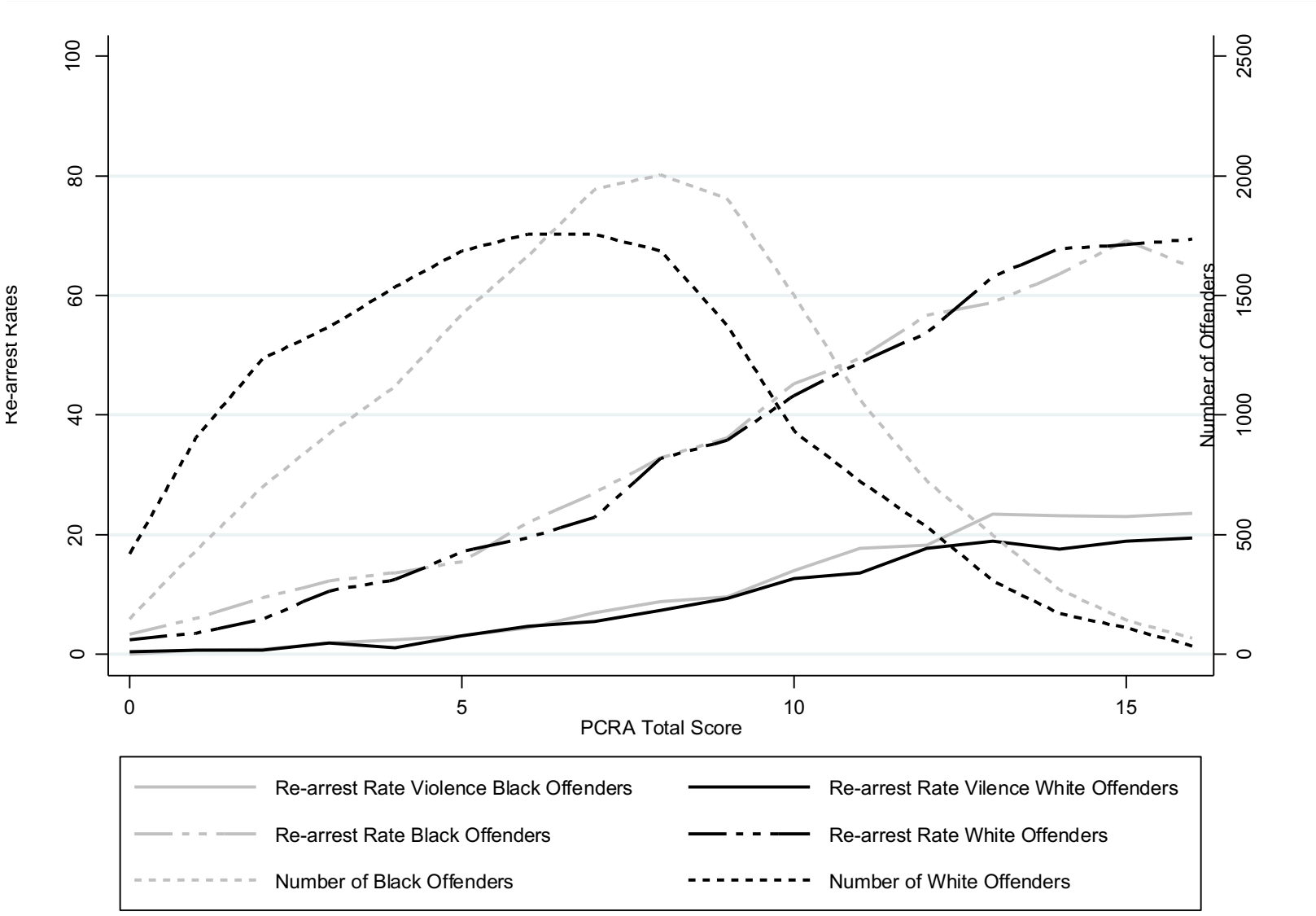
Electronic copy available at: https://ssrn.com/abstract=2687339

Figure 1. Predicted Probabilities of Arrest by PCRA Score and Race



Electronic copy available at: https://ssrn.com/abstract=2687339

Figure 2. Rate of Arrest and PCRA Distribution by Race



Electronic copy available at: https://ssrn.com/abstract=2687339

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, 2nd 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 .

Electronic copy available at: https://ssrn.com/abstract=2687339

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, 2nd ed.* New York: Plenum Press.

Electronic copy available at: https://ssrn.com/abstract=2687339

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.

Electronic copy available at: https://ssrn.com/abstract=2687339

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.

Electronic copy available at: https://ssrn.com/abstract=2687339

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

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

Electronic copy available at: https://ssrn.com/abstract=2687339

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

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, 4th ed. Downloaded 10/10/15 from: 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.

Electronic copy available at: https://ssrn.com/abstract=2687339

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, 5th 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.

Wrobleswki, J. (2014). *2014 US Department of Justice Criminal Division Annual Letter to US Sentencing Commission*

Electronic copy available at: https://ssrn.com/abstract=2687339

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

Electronic copy available at: https://ssrn.com/abstract=2687339

Endnotes

[i] Effect sizes were calculated by the first author based on data shared by Frase et al. (2015).

[ii] 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.

[iii] 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.

[iv] PCRA total scores greater than 16 were recoded to 16 as only 18 offenders have a PCRA total score of 17 or 18.

[v] 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.

Electronic copy available at: https://ssrn.com/abstract=2687339