DRAFT: November 5, 2015

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: 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. These instruments figure prominently in current reforms, but controversy has begun to swirl around their use. The principal concern is that benefits in crime control will be offset by costs in social justice—a disparate and adverse effect on racial minorities and the poor. Based on a sample of 34,794 federal offenders, we empirically examine the relationships among race (Black vs. White), actuarial risk assessment (the Post Conviction Risk Assessment [PCRA]), and re-arrest (for any/violent crime). First, application of well-established principles of psychological science revealed no real evidence of test bias for the PCRA—the instrument strongly predicts re-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 modestly higher average scores on the PCRA than White offenders (d= .43; appx. 27% non-overlap in groups' scores). So some applications of the PCRA could create disparate impact—which is defined by moral rather than empirical criteria. Third, most (69%) of the racial difference in PCRA scores is attributable to criminal history—which strongly predicts recidivism for both groups and is embedded in sentencing guidelines. Finally, criminal history is *not* a proxy for race—instead, it fully mediates the otherwise weak relationship between race and re-arrest. Data may be 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: 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 its 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 many risk factors that are sometimes included (e.g., marital history, employment status, neighborhood disadvantage) are

"proxies" for minority race and poverty (Starr, 2014; see also Harcourt, 2014; Silver & Miller, 2002). In the view of Former Attorney General Eric Holder (2014), risk assessment

> "may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society. Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant's history of criminal conduct. They should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place."

These concerns are legitimate and important—but untested. In fact, Holder specifically urged that this issue be studied. The main issue is whether the use of risk assessment in sentencing affects racial disparities in imprisonment, given that young black men are about 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 concerns are currently focused 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.). Whether the pivot point is at the front-end or back-end—and whether the decision is 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. Moreover, even the well-established use of risk assessment to inform resource allocation in corrections (see Casey,

Warren, & Elek, 2015) can invoke concern. If higher-risk offenders are subject to more intensive community supervision and risk reduction services (and service refusal violates the terms of release), they are more subject to social control than their lower-risk counterparts.

Does risk assessment exacerbate, mitigate, or have no effect on racial disparities? The answer to this question probably depends on factors that include the instrument chosen. Sensationalistic headlines aside, "risk assessment" is not reducible to "race assessment" (Sentencing Project, 2015). There are important differences among validated risk assessment instruments 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 direction for instrument selection and refinement. To provide context for this study, we first highlight where risk assessment fits in corrections and sentencing, and then unpack controversy about particular types of risk factors. After discussing how to evaluate test fairness, we present our specific aims.

Risk Assessment in (Community) Corrections

Risk assessment has been applied to inform correctional decisions for nearly a century (Administrative Office of the U.S. Courts, 2011; Andrews, Bonta & Wormith, 2006). Early instruments were designed to achieve efficient and effective prediction; they generally involved scoring a set of risk markers (e.g., young age, criminal history), weighting them by predictive strength, and combining them into a numerical score and/or a classification (e.g., low/medium/high risk). These classifications were 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 typically meant to scaffold efforts to implement evidence-based principles of 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 and his colleagues noted the potentially discriminatory effects of risk assessment in both juvenile- and criminal-justice settings, and illustrated how to remove "invidious predictors" (Gottfredson et al., 1994; Gottfredson & Jarjoura, 1996). Since then, little concern has been expressed about correctional applications of risk assessment. In fact, risk assessment and risk reduction play a central role in The Sentencing Reform and Corrections Act of 2015, a bill with bipartisan support before congress. This bill requires that risk assessments be conducted to assign federal inmates to appropriate recidivism reduction programs (e.g., work and education programs, drug rehabilitation). Eligible prisoners who successfully 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 are attracting 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 hyrbrid of retributive and utilitarian theories called "limiting retributivism" (Morris, 1974). As operationalized in the Model Penal Code (American Law Institute, 2014), sentencing takes place "within a range of severity proportionate to the gravity of offenses, [and] the blameworthiness of offenders." Within this range, a sentence is chosen to promote "offender rehabilitation [and] incapacitation of dangerous offenders" (§1.02(2), p. 2). That is, retributive concerns set a permissible range for the sentence (e.g., 5-9 years), and risk assessment is used to select a particular sentence within that range (e.g., 8 years for high risk). Risk assessment should never be used to sentence offenders to more time than they morally deserve.

Controversial Risk Factors

Risk factors irrelevant to blameworthiness (Starr's objection to socioeconomic factors). As explained by Monahan & Skeem (in press), 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. This tendency to conflate risk with blame fuels controversy about—and can constrain—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* marks increased likelihood for 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. Still more controversial factors (e.g., victimization) mitigate blameworthiness but increase risk.

According to an active critic of risk assessment (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, neighborhood, and financial background, which in her view are illegitimate—*both* because they are unrelated to moral culpability for past crime *and* because they are perceived as "proxies" for poverty and minority status. In Starr's arguments, blame seems to eclipse risk, as a concern appropriate to consider at sentencing. Criminal history—which increases perceived blameworthiness—may be the only risk factor she would call "in," for the purposes of sentencing.

Risk factors associated with race (Harcourt's objection to criminal history). In sharp contrast to Starr, another critic—Bernard Harcourt (2008)—categorically objects to the use of criminal history to inform sentencing, whether the vehicle is sentencing guidelines (which heavily 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."

There is evidence that minority race and criminal history are correlated (e.g., Durose, Snyder & Cooper, 2015)—but the degree varies as a function of how criminal history is operationalized. For example, in a meta-analysis of 21 studies that focused on racial differences on a measure often used to assess risk, Skeem, Edens, Camp & Colwell (2004) found negligible differences (d= .06) between Black and White groups on a multi-item criminal history scale (i.e., early conduct problems, juvenile delinquency, revocation of conditional release, poor anger

controls, criminal versatility) that robustly predicts recidivism (Walters, 2012). Moving from research to practice, Frase, Roberts, Hester, & Mitchell (2015) found wide variability in how 18 federal and state jurisdictions operationalized criminal history in their sentencing guidelines. Based on sentencing data from four jurisdictions, Frase et al. (2015) found that Black offenders obtained higher average scores than White offenders (*Mean d*= .24, *SD*=.05).[i] All four effect sizes were in the "small" range (*d*= .19-.29), suggesting about 79-85% overlap between Black and White groups in criminal history (see Cohen, 1988).

Of course, prior criminal convictions reflect 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 has long been hotly debated (see Frase 2014; McCord, Widom & Crowell, 2001), and varies as a function of crime type (e.g., violence vs. drug crimes) and stage of justice processing (e.g., arrest vs. incarceration decisions; see Blumstein et al., 1983; Piquero, 2015).

Risk factors that cannot be changed (Holder's objection to "static" characteristics). Starr (2015) suggests that risk factors "within the defendant's control" may legitimately be considered. 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 that are oriented toward the reduction of recidivism explicitly include variable risk factors—i.e., 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 effectively treated to reduce recidivism (see 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 through intervention. As is clear from this brief review, critics disagree in calling potentially race-related risk factors like criminal history "in" or "out," for the purposes of sentencing.

Bringing Psychological Science to the Controversy

Test bias vs. disparate impact. Data may be more helpful than rhetoric, if the goal is to improve sentencing and correctional practices at this opportune moment in history. Ample guidance on racial fairness in assessment is available from similar efforts undertaken in more mature fields (e.g., intelligence and other cognitive tests used to inform high-stakes education and employment decisions, see Reynolds 2000; Sackett, Borneman & Connelly, 2008). There is substantial agreement on the empirical criteria that indicate when a test is biased. These criteria have been distilled in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014)—which we refer to as the "Standards."

Given that the raison d'etre for risk assessment instruments is to predict recidivism, the paramount indicator of test bias is *predictive bias* (also known as "differential prediction;" Standard 3.7). On utilitarian grounds alone, any instrument used to inform sentencing must be shown to predict recidivism with similar accuracy across groups. If the instrument is unbiased, a given score will also have the same meaning regardless of group membership (e.g., an average risk score of X will relate to an average recidivism rate of Y for *both* Black and White groups). This is commonly tested by examining whether groups systematically deviate from a common regression line that relates test scores to the criterion (Cleary, 1968; see also Sackett & Bobko, 2010). In essence, statistical tests are performed to determine whether race moderates an instrument's predictive utility.

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, or vice versa). These instruments are not necessarily biased: "subgroup mean differences do not in and of themselves indicate lack of fairness" (The Standards, #3.6, p. 65). The notion that mean differences are indicative of test bias is unequivocally rejected in the professional literature because group differences in scores may reflect true differences in recidivism risk, based on group variation "in experience, in opportunity, or in interest in a particular domain" (Sacket et al., 2008, p. 222). Race reflects to deep and 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. For this reason, the Standards (3.6) suggest that instruments be examined to understand and (if possible) reduce group differences. Similarly, if two instruments are equally valid "and impose similar costs," the Standards (3.20) advise "selecting the test that minimizes subgroup differences." Fundamentally, disparate impact is a moral consideration. As Frase (2013) has noted, 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).

In our view, risk assessment instruments used at sentencing—and the risk factors they subsume—must be empirically examined for both predictive bias (moderation by race) and disparate impact (association with race). Simply put, risk assessment must be both empirically valid and perceived as morally fair across groups.

This study is among the first to rigorously examine the relations among risk, race, and recidivism among adult offenders in the U.S. Although this issue has been studied with juvenile offenders (e.g., Olver et al., 2009), forensic instruments designed to predict violence (e.g., Singh & Fazel, 2010), and indigenous/non-indigenous groups in other countries (e.g., Wilson & Gutierrez, 2014), our focus is on comparing Black and White offenders in the U.S. on

instruments designed to predict recidivism. In a recent meta-analysis, Desmarais, Johnson, & Singh (in press) identified 53 published and unpublished studies of 19 risk assessment instruments used in U.S. correctional settings. In keeping with rigorous meta-analyses (e.g., Yang, Wong, & Coid, 2010), the authors found that the predictive accuracy of these instruments was essentially interchangeable. More to the point, although only three studies permitted comparisons of predictive accuracy by offender race/ethnicity (Desmarais et al., in press), results indicated that levels of predictive utility for Black and White offenders were identical (AUCs=.69 on the "COMPAS;" Brennan et al., 2009) or highly similar (ORs=1.03 [Black] and 1.04 [White] on the Levels of Services Inventory-Revised or LSI-R; Lowenkamp & Bechtel, 2007; Kim, 2010). Formal tests of predictive bias (moderation) 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. Critics of risk assessment often use the term "proxy" to refer to some risk factors—or total risk scores (see Harcourt, 2015; Starr, 2014). The term suggests that these factors are so highly correlated with poverty or minority race that they can be used as indirect indicators to "stand in" for suspect variables that are not measured directly. Often, however, it is not clear that factors like criminal history are meant to proxy for race (i.e., are meant to intentionally camouflage discrimination).

Progress is possible when terms like "proxy" are operationally defined. Kraemer et al. (2001) clarify how risk factors can work together to predict an outcome like recidivism. In their terminology, a proxy is a correlate of a strongly predictive risk factor that also appears to be a risk factor for the same outcome—but the only connection between the correlate and the outcome is the strong risk factor correlated with both. By their criteria, criminal history is a

proxy for race only if race "dominates" in predicting recidivism (i.e., maximum strength in predicting recidivism is achieved by race alone – not criminal history alone; not the combination of criminal history and race). This is highly unlikely. Criminal history typically predicts recidivism much more strongly than race, so will dominate or codominate race (Berk 2009; Durose, Cooper & Snyder 2014). If so, criminal history is not a proxy for race—instead, it overlaps race and possibly mediates race's relation to recidivism (i.e., criminal history 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 as a whole, risk assessment is not used to inform front-end sentencing decisions. Instead, the Federal Post Conviction Risk Assessment or "PCRA" (Johnson, Lowenkamp & VanBenschoten, 2011) is administered post-conviction, upon intake to a term of supervised release or probation in the community. The PCRA is used to inform decisions designed to reduce risk—i.e., to identify *whom* to provide with the most intensive supervision and services (i.e., higher-risk offenders, leaving lower-risk offenders with less intensive versions) and *what* to target in those services (i.e., variable risk factors). The PCRA was developed by the Administrative Office of the US Courts (Probation and Pretrial Services Office) to improve the effectiveness and efficiency of post-conviction community supervision—and should not be used for other sentencing 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. Thus, these federal data are well-suited for addressing four aims with broader implications:

1. To what extent is the instrument—and the risk factors it includes—free of *predictive bias?* We hypothesize that there will be little or no evidence that the accuracy of the PCRA in predicting re-arrest depends on whether offenders are Black or White.
2. To what extent does the instrument yield average score differences between racial groups that are relevant to *disparate impact*? We hypothesize that Black offenders will obtain similar—or modestly higher—PCRA scores than Whites, given past research.
3. Which risk factors contribute the most and the least to mean score differences between Black and White offenders? Given past research, we expect criminal history to contribute the most to these differences—and variable risk factors like substance abuse to contribute the least.
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

Participants in this study were drawn from a larger dataset on over 96,000 offenders who were assessed 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 = 29,680), (b) no missing data on PCRA items (to permit analyses at the risk factor level; *n* lost = 1,007), and (c)

race coded as either "Black" or non-Hispanic "White" (to permit relevant racial comparisons*; n* lost = 17,238). Application of these criteria yielded an eligible pool of 48,475 offenders.

In the eligible pool, Black participants were more likely than Whites to be male and young—and both characteristics are risk factors for recidivism. To more precisely estimate the effect of race on risk (especially for mean score comparisons), we randomly matched each Black offender to a White offender on age and sex, using ccmatch in STATA (Cook, 2015).[ii] This process yielded a sample of 34,794 offenders—17,397 Black and 17,397 White. Compared to the larger sample from which it was drawn (Walters & Lowenkamp, 2015), the present study sample is similar in age, sex, conviction offense, and PCRA total scores.

Sample characteristics are shown in Table 1. Because 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. As shown in Table 1, offenders' average age was 39 and the vast majority were male. For both Black and White offenders, the modal conviction offense was for a drug crime. Although all offenders were followed for a minimum of one year, the follow up period (i.e., time at risk for re-offending) was variable beyond that point. There were no significant differences between Black and White offenders in their average follow up time ($t(34744.5) = -2.81$; $p = 0.005$).

[Insert Table 1]

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 (Johnson et al., 2011). Fifteen items are scored and summed to yield a total PCRA risk score 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 considered 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
- "Employment and education" includes highest grade completed, unstable recent work history, and currently unemployed
- "Social networks" includes family problems, unmarried, and lack of social support
- "Substance abuse" includes recent alcohol problems and recent drug problems
- "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 officers when an offender entered supervision or when reassessing an offender. In the present study, the results of the earliest assessment were selected for analyses as this provided the longest follow up time period. In addition to the total PCRA score, the sub-scores from the PCRA domains (criminal history, education & employment, drugs & alcohol, social networks, and cognitions) were also calculated and used in some analyses.

Arrest Criterion

Data from the National Crime Information Center (NCIC) and Access to Law Enforcement System were used to collect information on arrests. A standard criminal history check was retrieved on each participant that yielded their entire criminal history. The date and types of arrests that occurred after the date of PCRA administration were coded from these data. The result was two dichotomous measures that we used in analyses of predictive fairness: arrest for any offense (excluding technical violations of standard conditions of supervision), and arrest for any violent offense. Violence was defined using the NCIC definitions (i.e., homicide and related offenses, kidnapping, rape and sexual assault, robbery, assault).

To promote clarity and reading ease, our analyses primarily focus on "any arrest." We also report analyses for "violent arrests," given the importance of using a valid criterion for assessing predictive fairness. 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).

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 29% (32% Black; 25% White, $\chi^2(1) = 261.35$; $p < 0.001$; $\phi = -0.09$), and the base rate for violent arrest was 8% (9% Black; 6% White, $\chi^2(1) = 127.66$; $p < 0.001$, $\phi = -0.06$). Black participants were modestly more likely to be arrested than Whites.

Analyses

To address our aims, we calculated descriptive statistics, effect sizes (Cohen's *d*), and measures of predictive validity (AUCs and DIF-R; Silver, Smith & Banks, 2000). We also performed regressions to test whether race moderated the predictive utility of the PCRA and to classify risk factors as mediators or proxies, according to Kraemer et al's (2001) criteria.

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.

Degree of prediction, as a function of race. First, we examined whether the *degree* of relationship between PCRA total scores and re-arrest varied as a function of race (see Arnold, 1982). Table 2 presents re-arrest rates for offenders classified in each PCRA risk classification (low to high) by race. Note that re-arrest rates increase monotonically as risk classifications increase, across racial groups—and that re-arrest rates within a given classification are fairly similar, across racial groups.

[Insert Table 2]

Ideally, risk classifications create reasonably sized groups of offenders with re-arrest rates that are as different as possible. We used the Dispersion Index for Risk (DIFR; see Silver, Smith & Banks 2000) to test the extent to which the PCRA maximizes "base rate dispersion" for Black and White groups. DIFR ranges from 0 to infinity, increasing as the classification model disperses cases into subgroups whose baserates of re-arrest are distant from the total sample baserate and whose subgroup sample sizes are large in proportion to the total sample size. As shown in Table 2, PCRA classifications performed somewhat better for White (DIFR= 1.10 & 1.09 for "Any" and "Violent," respectively) than Black (DIFR=0.84 & 0.91 for "Any" and "Violent," respectively) participants. Because no formulae are available to estimate confidence intervals for the DIFR, the significance of this difference is unclear. However, DIFR values for both Black and White groups are high, compared to other risk assessment tools implemented in "real world" settings (see Skeem et al., 2013—DIFR=.68-.71 for tools that performed relatively well). At an absolute level, the PCRA seems to adequately classify Black and White offenders.[iii]

Risk classifications serve important functions in practice, but are less precise than total scores. To test the predictive utility of PCRA total scores by race, we used a measure of association called the Area Under the ROC Curve. The AUC is widely used in this context, partly because its values are not heavily influenced by base rates of offending (which vary across samples and studies). 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, the AUC values for Any Rearrest (Black=.73, White=.77) and for Violent Rearrest (Black=.73, White=.76) are large, across race. The latter comparison indicates a 73% (Black) or 76% 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 difference in predictive utility across groups is small (AUC_{diff}=.03-.04), and reached statistical significance only for "any" rearrest.

Form of prediction as a function of race. Having found that PCRA scores account for roughly the same degree of variance in re-arrest among Black offenders as White offenders, we next examined whether the *form* of the relationship between PCRA scores and recidivism varies as a function of race (Arnold, 1982). If the instrument is unbiased, a change in a PCRA score will make the same amount of difference in re-arrest rates for Black as White offenders.

To test whether race moderates the relationship between the PCRA and re-arrest, we estimated a series of logistic regression models. As shown in Table 3, the first model included only race; the second model included only PCRA scores; and the third model included both race and PCRA scores. The fourth model added the interaction term of interest between race and PCRA scores. A comparison of Models 3 and 4 reveal no significant difference in the χ^2 fit of these models and no change in pseudo R^2—indicating that the addition of the interaction term

does not improve the prediction of re-arrest. In fact, a comparison of Bayesian Information Criterion (BIC) values between Models 2, 3, & 4 (BIC = 35742.5, 35749.6, & 35750.1) strongly favors Model 2 (PCRA only) over models that include race. Moreover, the odds ratio for the interaction term in Model 4 is trivial (1.03) and not statistically significant. Similar results were obtained for parallel analyses that used "violent" rearrest as the criterion (Models 3 & 4 had similar χ^2 values; the interaction term was trivial [OR=1.01], and not statistically significant).

[Insert Table 3]

Recall that participants' length of follow-up varied. To account for variable time at risk, we also tested for moderation by completing a series of four Cox survival analyses that parallel the regression models described above. The pattern of results was similar, both for "any" re-arrest and "violent" rearrest: That is, Models 3 & 4 had similar χ^2 values, the interaction term was very small [HR=1.04 & 1.02,for "any" and "violent," respectively], and reached statistical significance only for "any" rearrest.

Again, trivial differences can become statistically significant in samples as large as ours (Lin et al., 2013). To concretize any racial differences in the form of the relation between the PCRA and re-arrest, we (a) estimated the predicted probabilities of any re-arrest based on moderated regression Model 4 reported above (see Table 3),[iv] (b) grouped those probabilities together for each PCRA score,[v] and (c) displayed those grouped probabilities by race in Figure 1. Given the trivial odds ratio, one would expect—and one observes—that the two lines would be nearly identical. Across PCRA scores, the predicted probabilities of re-arrest for Black and White offenders are much more similar than dissimilar in form (elevation and shape).

[Insert Figure 1]

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 (i.e., point biserial correlations) between PCRA domain scores and re-arrest, by race. As shown there, criminal history had a medium effect in predicting re-arrest, and the remaining four domains had a small effect. Although substance use and social networks predicted statistically significantly better for White than Black participants, there were no racial differences across the other three domains.

[Insert Table 4]

Next, we assessed whether race moderated the relation between PCRA risk factors and any re-arrest. For each risk domain, we completed a series of four logistic regression models that parallel those described above for PCRA total scores. Table 5 displays each of the five risk domains (in Column 1), the change in pseudo-R^2 (Column 2) and step χ^2 (Column 3) when the interaction term (risk domain x race) was added to the main effects model, and the odds ratio for the interaction term (in Column 4). Results were consistent with those reported above. Specifically, race moderated the effect of substance use and social networks—but criminal history, employment and education, and attitudes predicted re-arrest similarly for Black and White participants.

[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, there was evidence that the substance abuse and social network domains predicted any re-arrest better for White than Black offenders. There is no evidence of predictive bias for the remaining risk domains—including criminal history and employment and education.

Assessing Mean Score Differences Relevant to Disparate Impact

The second aim was to assess the extent to which the PCRA yields average score differences between racial groups relevant to *disparate impact*. We hypothesized that Black offenders would obtain similar—or modestly higher—PCRA total scores than Whites.

The mean PCRA total score was 7.33 (*sd*= 3.20) for Black participants and 5.93 (*sd*= 3.40) for White participants—an average 1.4 point difference on an 18-point scale. According to conventional classifications, minimum *d* values of .2, .5, and .8 define small, medium, and large effects, respectively (Cohen, 1988). By these standards, the effect of race on PCRA scores—i.e., *d*= .43 (95% CI=.41-.45)—is "small." A *d* of .40 corresponds to 73% overlap (and 27% non-overlap) between Black and White groups in PCRA scores (see Cohen, 1988). So the difference in scores between groups is small—but potentially meaningful.

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 to these differences—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 Table 6, along with Cohen's *d*.

Column 8 indicates the percentage of the difference in the PCRA total means that is attributable to a given risk domain. As shown in that column, 69% of the racial difference in mean PCRA scores is attributable to differences in criminal history (which differ an average of 0.97 point on a 9-point scale). In fact, the effect of race on criminal history (*d*= .43; *CI*= .41-.45) is the same as that of total PCRA scores.

[Insert Table 6]

Most of the remaining racial difference in average PCRA total scores—i.e., 24%--is attributable to the employment and education domain (which differs an average of 0.34 points on a 3-point scale). The effect size is "small" (*d* = 0.36, *CI* = .34-.38). Again, this is the PCRA domain that manifests the most change over time (Cohen & VanBenschoten, 2014).

The remaining three PCRA domains—substance abuse, attitudes, and social networks—contributed negligibly (a total of 7%) to mean score differences between Black and White offenders. Effect sizes for these domains tended to fall near or below *d*=.10, which corresponds to 92% overlap between Black and White groups (Cohen, 1988).

Drilling down on criminal history. Criminal history can be measured in myriad ways that are more- or less- correlated with race. For this reason, Frase et al. (2015) recommended that individual items be examined by race. In Table 7, we display mean score differences by race for five of the six criminal history items. The sixth item—age—is omitted because the sample was matched on age and sex to isolate differences specific to race (in the eligible pool, Black offenders were modestly more likely to be young than White offenders, *d*= .34 [*CI*=0.33-0.36]). As shown in Table 7, the effect of race for each criminal history item is "small," 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.

[Insert Table 7]

Given the importance of criminal history to mean score differences on the PCRA, we also explored whether race moderated the utility of each item in predicting any arrest (i.e., tested for predictive bias at the item level). Briefly, results suggest that race moderates the effect of age in predicting recidivism, with age predicting re-arrest more strongly for Black than White offenders (details available from authors). Race did not moderate the utility of the remaining five criminal history items in predicting arrest.

Proxy or Mediator?

The final aim was to examine whether variables like criminal history are best understood as proxies for race or mediators of the relation between race and recidivism. We expected the best classification would be "mediator."

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 rearrest?). Applying these criteria, race precedes criminal history (by definition) and race and criminal history are correlated (d=.43; see Table 5). The issue is whether race "dominates" in predicting any arrest. It does not—intead, criminal history (r_{pb} = .35) predicted arrest more strongly than race (ϕ =-.09). So criminal history is not a proxy for race—instead, it overlaps race and mediates the relation between race and recidivism.

Does criminal history *partially* or *totally* explain the relation between race and recidivism? Put in Kraemer et al.'s parlance, does criminal history alone predict recidivism most strongly (i.e., criminal history dominates race) or do criminal history and race together predict

recidivism most strongly (i.e., the variables codominate)? To address these questions, we completed a series of mediation analyses using the binary_mediation package (Ender, 2011). As shown in Table 8, we found that (a) the mediating variable (criminal history) was associated with the primary explanatory variable (race; see Panel 1), (b) the primary explanatory variable (race) predicted the outcome of interest (any arrest; see Panel 2), and (c) the relationship between the primary explanatory variable (race) and the outcome of interest (any arrest) was totally mediated by criminal history (see Panel 3). As shown in lower panel, 85% of the effect of race on re-arrest was mediated by criminal history.[vi]

[Insert Table 8]

Putting Predictive Fairness and Mean Score Differences Together

Figure 2 provides a visual summary of the study's global findings (adapted from Monahan et al., 2001). In this figure, PCRA scores appear on the X axis and percentages (0-100%) appear on the Y axis. We plotted the percentage of the group with each PCRA score that is Black as a line, along with re-arrest rates for any crime by race as a bar chart. Recall that 50% of the sample is Black and 50% is White. This figure shows that (a) the percentage of offenders who are Black increases—at least to the mid-point of the scale—as PCRA scores increase—i.e., the positive slope of the line shows the small effect of race on total scores, d= .43, and (b) actual re-arrest rates for both White and Black offenders (shown in the bar chart) increase steeply and similarly as PCRA scores increase, in line with the predicted probabilities shown in Figure 1.

In Figure 3, a more dimensional visual summary is provided—one that includes re-arrest for a violent crime as an outcome. In this figure, PCRA scores appear on the X axis and percentages (0-100%) appear on the left Y axis while the number of offenders (0-2,000) appear on the right Y axis. We plotted the re-arrest rates for any crime and violent crime by race against

the left Y axis—and the number of Black and White offenders with each PCRA score against the left Y axis. This figure shows (a) the small (estimated 27%; see above) non-overlap between Black and White groups in PCRA distributions—much of it falling at the lower end of the scale, and (b) the steep and similar increase in re-arrest rates for Black and White offenders for both any arrest (depicted earlier) and violent arrest (depicted only here).

DISCUSSION

At the most basic level, these results indicate that risk assessment is not "race assessment." First, there is no real 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. The difference is small (d= .43), but potentially meaningful (≈27% non-overlap in scores). So some applications of the PCRA could create disparate impact—which is defined by moral rather than empirical criteria. Third, most (69%) 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 (nor is risk itself a proxy for race). Instead, criminal history fully mediates the otherwise weak relationship between race and re-arrest.

Are these results merely a function of "bias predicting bias," e.g., biased criminal history records predicting biased future police decisions about arrest? Put more broadly, is the appearance of validity for the PCRA due to differential selection? In a word—no. First, criminal history predicts arrest with the same strength and form, whether participants are Black

or White. And neither criminal history nor risk as a whole function as "proxies" for race. 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), PCRA "need" scores (i.e., employment-education, social networks, substance abuse, and attitudes; *OR* = 1.18, p<.001) add significant incremental utility in predicting arrests for violence for both Black and White participants, $\Delta\chi^2$ (1) = 204.67, 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 each finding, though, we note two study limitations that must be borne in mind. First, we used a sample of Black and White offenders matched in age and gender. Because this study is among the first to focus on the topic, we wished to isolate the effects of race on risk and recidivism. Parallel analyses completed with the eligible (non-matched) sample yielded a similar pattern of results—which lends confidence our findings. Given that race can interact with other characteristics (e.g., young age increases the likelihood of rearrest more for Black than White offenders) in a manner relevant to sentencing disparities (e.g., Steffenmeier, Ulmer, & Kraemer, 1998), we recommend that more complex models be examined in future work. Second, data on interrater reliability in scoring the PCRA are not available for the present sample. Although some risk domains and/or groups may have been scored more accurately than others, all officers who complete the PCRA must complete a certification process that has been shown to yield reliable scores (Lowenkamp et al., 2013).

Lack of Test Bias

The degree and form of association between PCRA total scores and re-arrest were essentially the same, for Black and White offenders. These findings are consistent with past studies indicating that the degree of association between other "risk-needs" tools (i.e., the LSI-R and COMPAS) and recidivism are similar for Black and White offenders (Brennan et al., 2009; Lowenkamp & Bechtel, 2007; Kim, 2010). This research goes beyond past findings by testing whether the form of the relationship between risk and recidivism is similar across races. We found that race did not moderate the utility of the PCRA in predicting a new arrest for a violent crime, even when time at risk for re-rearrest was taken into account. According to principles that have been well-established in the arena of high stakes testing, a given score must be shown to have the same meaning, regardless of group membership—as we have done here.

The most appropriate level for assessing test fairness is the test level—rather than the scale level. However, having established a lack of predictive bias for PCRA total scores, we also examined predictive bias at the level of specific risk factors because (a) the results are relevant to other instruments with similar risk factors, and (b) some factors—especially criminal history and employment and education—have been labeled as racially unfair by critics (e.g., Harcourt, 2015; Starr, 2014). For three of the five risk domains—including those claimed to be racially unfair—their degree and form of relationship to re-arrest was the same for Black and White offenders. Predictive bias was evident for only two factors—i.e., recent substance abuse problems and social networks (i.e., unmarried, family problems, lack of social support), which predicted modestly more strongly for White than Black offenders (see Table 4).

As these results imply, risk assessment instruments that are very short and/or have been developed with fairly homogeneous samples may be more prone to predictive bias than instrument examined here. The utility of particular risk factors in predicting recidivism can

differ across groups (for differences by developmental stage, see Herrenkohl et al., 2000). Moreover, definitions of particular risk constructs may not completely overlap across groups; behaviors relevant to the construct may be poorly sampled, or there may be incomplete coverage of all facets of the construct (e.g., "unmarried" may be less indicative of the "social network problem" construct for Black than White offenders, see Bureau of Labor Statistics, 2013; van de Vijver & Tanzer, 2004). Risk assessment instruments with broad coverage that are developed with diverse samples may include predictive items that distinguish some groups from others. This may not be bad, from a psychometric point of view. In fact, in tests of measurement bias in the cognitive testing literature, it is "common to find roughly equal numbers of differentially functioning items favoring each subgroup, resulting in no systematic bias at the test level" (Society for Industrial and Organizational Psychology [SIOP], 2003, p. 34).

In short, despite limited evidence of predictive bias at the risk factor level, we found no evidence of test bias for the PCRA itself. Scores on the PCRA are useful for forward-looking assessments of an offender's risk of future crime, whether the offender is Black or White. The generalizability of these results to other risk assessment instruments is unclear. Because instruments differ in their breadth of content and quality of development, tests of predictive bias should be routinely conducted (see The Standards, 3.7).

Mean Score Differences Relevant to Disparate Impact

Small but potentially meaningful differences. Mean score differences between groups are uniformly rejected as an indicator of test bias because group differences may reflect real differences (e.g., 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 "disparities" are a salient issue, given that Black offenders are already incarcerated at a much greater rate than White offenders.

We found a "small" effect of race on PCRA scores—i.e., an average group difference of 1.41-points in total scores, d= 0.43, roughly corresponding to 27% non-overlap between Black and White groups (Cohen, 1988). This is similar to the "small" effect of race on criminal history scores that are embedded in sentencing guidelines (d= .19-.29; or 15-21% non-overlap; data from Frase et al., 2015). For the sake of broader comparison, these effects pale in comparison to those observed for high stakes cognitive tests. The results of a comprehensive meta-analysis indicate a "large" effect of race on these tests, including the SAT (d =0.99), ACT (d=1.02) and GRE (d= 1.34; Roth, Bevier, Bobko, Switzer & Tyler, 2001). These effect sizes roughly correspond to a 55-65% non-overlap between Black and White groups (Cohen, 1988).

When are mean score differences large enough to translate into disparate impact? There are no set criteria for addressing this question because disparate impact is defined by moral concerns. 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).

Disparate impact is about the *use* of the instrument (not the instrument itself). 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 offenders might be slightly 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 about social control—more surveillance and more conditions of supervised release—not service access (see Swanson et al.,

2009). Of course, this latter view must be juxtaposed against an established tradition of relying upon risk assessment as a factor in probation, parole, and other accelerated release practices designed to use correctional resources efficiently while protecting public safety.

In an effort to begin addressing these nebulous issues, 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 Black and White offenders. Criminal history accounts for most (69%) of the difference in PCRA scores (d=.43)—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). Within the criminal history domain, the item "number of past offenses" accounts for almost half (45%) of domain differences—mostly because this item is weighed more heavily than other items (e.g., violent offenses) with similar effect sizes (all "small"). This finding is consistent with Frase et al.'s (2015) observation that Black and White offenders systematically manifest small differences in criminal history scores, with the magnitude varying as a function of how sentencing guidelines operationalize this variable.

Criminal history presents a conundrum. On one hand, criminal history is among the strongest predictors of (violent) re-arrest—for both Black and White offenders (see Table 4). And—compared to other risk factors, criminal history is more relevant to an offender's perceived

blameworthiness for the conviction offense (Monahan & Skeem, in press). This may help explain why criminal history has quietly become embedded in many jurisdictions' sentencing guidelines, unlike risk factors that do not bear on an offender's blameworthiness (e.g., education and employment). On the other hand, heavy reliance on criminal history at sentencing (whether in the form of sentencing guidelines or risk assessment) will contribute more to disparities in incarceration than reliance upon other robust risk factors that are less bound to race.

These concerns about criminal history are loosely consistent with Harcourt's (2015) criticisms. However, criminal history is not a proxy for race (as Harcourt contends)—it is not the case that the principal connection between criminal history and re-arrest is race. Criminal history is better construed as a mediator. There is a trivial relationship between race and re-arrest; a small relationship between race and criminal history; and a moderate relationship between criminal history and re-arrest. Although causal relationships cannot be inferred from non-experimental data, our results are consistent with what we would expect to see if a causal path leading from race to criminal history to re-arrest were in force (Kraemer et al., 2001).

The results of this study are less consistent with Starr's (2014) objections to risk assessment. The employment and education domain was free of predictive bias, manifested small mean score differences between Black and White offenders (d= .36), and accounted for only 24% of the small difference in PCRA total scores. Moreover, employment and education scores—at least operationalized in the PCRA—have been found to change over relatively short periods of time: Among high-risk offenders, for example, 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 trivial (*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 directly relevant to risk reduction efforts (see Monahan & Skeem, in press).

Familiar dilemma. In summary, the PCRA—including the controversial domains of criminal history and employment and education—is free of predictive bias. Nevertheless, there are small mean score differences between Black and White offenders that could be meaningful, in terms of disparate impact, if the instrument were applied to inform sentencing.

The dilemma about predictive utility and disparate impact has long been familiar in the high stakes cognitive testing domain, where mean score differences between Black and White groups are <u>much</u> larger (see above) than those observed here. As summarized by Sackett et al. (2008, p. 222):

> "Particularly with regard to race and ethnicity, the [large] 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."

Here, the issue is that risk assessment instruments could scaffold contemporary efforts to unwind mass incarceration without compromising public safety. These instruments are directly

relevant to utilitarian goals of sentencing. But using some instruments in this manner might exacerbate existing racial disparities in incarceration. If one values one concern—predictive accuracy or social justice—to the exclusion of the other, there is no dilemma. If one values both concerns, which is likely to be the case most of the time, the goal is to balance the two goals (see Sackett et al., 2001).

Implications

The most straightforward implication of the present study is that risk assessment instruments should be routinely tested for predictive bias and mean score differences by race. For obvious reasons, these are fundamental standards of testing—particularly in high stakes domains (see The Standards, Section 3). We recommend that these issues be examined not only at the test level, but also at the level of risk factors. If policymakers blindly eradicate risk factors from a tool because they are contentious, they risk reducing predictive utility *and* exacerbating the racial disparities they seek to ameliorate. It may be politically tempting, for example, to focus an instrument tightly on criminal history because this variable is associated with perceptions of blameworthiness, and is also easily assessed by referring to conviction records. But risk estimates based on a broader set of factors predict recidivism better than criminal history and tend to be less correlated with race (e.g., Berk 2009).

As suggested above, a number of strategies have been tested for maximizing an instrument's predictive utility while minimizing mean score differences. For example, in the context of selection for employment and education, efforts have been made to identify other predictors of work- and academic- performance (e.g., personality, interests, socioemotional skills; Sackett et al., 2001). Reasoning by analogy, efforts could be undertaken in the risk assessment domain to rely less heavily on criminal history while weighting risk factors with

fewer mean score differences more heavily. Whether and how this strategy would "work" is unclear—and also beyond the scope of the present article (see Lowenkamp, Skeem, & Monahan, in preparation).[vii]

Conclusion

In light of our results, it seems that concerns expressed about risk assessment are exaggerated. To be clear, we are not offering a blanket endorsement of the use of risk assessment instruments to inform sentencing. There will always be bad instruments (e.g., tests that are poorly validated) and good instruments "used inappropriately (e.g., tests with strong validity evidence for one type of usage put to a different use for which there is no supporting evidence)" (Sackett et al., 2008, p. 225). We are simply offering a framework for examining important concerns related to race, risk assessment, and recidivism. Our results demonstrate that risk assessment instruments *can* be free of predictive bias and *can* be associated with small mean score differences by race. They also provide some direction for improving instruments in a manner that might balance concerns about predictive utility and disparate impact.

This article focuses on one factor that would influence whether the use of risk assessment in sentencing would exacerbate, mitigate, or have no effect on racial disparities in imprisonment—the instrument itself. But the instrument is only part of the equation. Given findings in the general sentencing literature, the effect of risk assessment on disparities will also vary as a function of the baseline sentencing context: Risk assessment, compared to what? Racial disparities depend on where one is sentenced (Ullmer 2012), so—holding all else constant—the effect of a given instrument on disparities will depend on what practices are being replaced (Monahan & Skeem, in press; see also Ryan & Ployhart, 2014).

Although practices vary, common denominators include (a) judges' intuitive and informal 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 has been at least one demonstration that risk assessment can be introduced without causing more punitive sentences for high-risk offenders (albeit in the Netherlands; see van Wingerden, van Wilsem, & Moerings, 2014). There is no empirical basis for assuming that the status quo—across contexts—is preferable to judicious application of a well-validated and unbiased risk assessment instrument. We hope the field proceeds with due caution.

Table 1: Sample Characteristics

Characteristic	All	Black	White
Age	39.18 (10.29)	39.18	39.18
% Male	84%	84%	84%
% Conviction offense^			
Drug	46	53	40
Firearms	15	16	14
White Collar	17	15	19
Public Order	6	5	8
Property	5	4	6
Violence	5	5	5
Sex offense	3	1	5
Average follow-up period in days	1035 (238)	1032 (243)	1039 (234)

^ Categories with less than 5% excluded

Table 2. Predictive Utility of PCRA Risk Classifications and Total Scores by Race

Feature	All	Black	White	All	Black	White
	Rearrest			Violent Rearrest		
% Rearrested by PCRA Classification						
Low	11	13	10	2	2	2
Low/Moderate	30	31	29	8	8	7
Moderate	54	55	53	17	18	16
High	73	71	77	24	25	21
DIF-R, PCRA Categories	0.90	0.84	1.10	1.04	0.91	1.09
AUC, PCRA Total[1]	0.75	0.73	0.77	0.75	0.73	0.76

[1]Difference is significant for Rearrest ($Z = -6.4284$; $p < 0.001$), but not for Violent Rearrest

Table 3. Logistic Regression Models Testing Whether Race Moderates the Utility of PCRA Total Scores in Predicting Rearrest

	Model 1	Model 2	Model 3	Model 4
White	0.68	--	0.95	0.77
PCRA Total	--	1.35	1.35	1.33
White X PCRA Total Interaction	--	--	--	1.03
Constant	0.48	0.04	0.04	0.05

Note: Values are odds ratios for each predictor. No terms were significant at $p < .001$
Model 1 Log Likelihood = -20701.55; $X^2(2) = 261.97$; $p < 0.001$; pseudo $R^2 = 0.01$; n = 34,794
Model 2 Log Likelihood = -17860.79; $X^2(2) = 5943.49$; $p < 0.001$; pseudo $R^2 = 0.14$; n = 34,794
Model 3 Log Likelihood = -17859.13; $X^2(3) = 5946.80$; $p < 0.001$; pseudo $R^2 = 0.14$; n = 34,794
Model 4 Log Likelihood = -17854.14; $X^2(4) = 5956.78$; $p < 0.001$; pseudo $R^2 = 0.14$; n = 34,794

Table 4. Correlation Between PCRA Domain Scores and Future Re-arrest

Risk Domain	All	Black	White	*Z*
Criminal History	0.35	0.33	0.37	2.11
Employment	0.23	0.22	0.23	-0.98
Substance Use	0.21	0.18	0.25	-6.85*
Social Networks	0.20	0.18	0.22	-3.89*
Attitude	0.16	0.16	0.15	0.96

* $p \leq .001$

Table 5. Results of Logistic Regression Models Testing Whether Race Moderates the Utility of Each PCRA Domain (Independently)

	Change in R^2	Step Chi Square	Odds Ratio For Interaction Term
Criminal History	0.001	7.59	1.04
Employment	0.001	9.48	1.08
Substance Use	0.001	34.80	1.31*
Social Networks	0.000	19.42	1.15*
Attitudes	0.000	2.35	1.12

*p< .001 for both step and interaction term

Table 6. PCRA Total and Domain Scores by Race

	Black			White					d		
Variable	N	Mean	Std. Dev.	N	Mean	Std. Dev.	Difference	% Attributable To	Estimate	Lower	Upper
PCRA Total	17,397	7.33	3.20	17,397	5.93	3.40	1.41		0.43	0.41	0.45
Criminal History	17,397	4.76	2.15	17,397	3.79	2.33	0.97	69	0.43	0.41	0.45
Employment/Education	17,397	1.14	1.01	17,397	0.80	0.90	0.34	24	0.36	0.34	0.38
Substance Abuse	17,397	0.21	0.48	17,397	0.22	0.50	-0.02	-1	-0.04	-0.06	-0.02
Social Networks	17,397	1.11	0.78	17,397	1.02	0.79	0.08	6	0.11	0.09	0.13
Attitudes	17,397	0.12	0.33	17,397	0.09	0.29	0.03	2	0.10	0.08	0.12

Table 7. PCRA Criminal History Item Scores by Race

	Black			White					d		
Variable	N	Mean	Std. Dev.	N	Mean	Std. Dev.	Difference	% Attributable To	Estimate	Lower	Upper
Prior Arrests	17,397	2.02	1.02	17,397	1.59	1.11	0.43	45	0.41	0.39	0.43
Violent Offenses	17,397	0.54	0.50	17,397	0.36	0.48	0.18	19	0.36	0.34	0.38
Varied Offending Pattern	17,397	0.77	0.42	17,397	0.63	0.48	0.14	14	0.31	0.28	0.33
CS Violation	17,397	0.48	0.50	17,397	0.36	0.48	0.12	12	0.25	0.23	0.27
Institutional Adjustment	17,397	0.26	0.44	17,397	0.17	0.38	0.09	9	0.23	0.21	0.25

Table 8. Mediation Analysis Criminal History Domain Score

OLS Model Predicting Criminal History Score				
	Coefficient	SE	t	p value
White	-0.96	0.02	-40.36	<0.001
Constant	4.76	0.02	280.38	<0.001
Logistic Regression Model Predicting Recidivism				
	Coefficient	SE	z	p value
White	-0.39	0.02	-16.13	<0.001
Constant	-0.73	0.02	-45.05	<0.001
Logistic Regression Model Predicting Recidivism				
	Coefficient	SE	z	p value
Criminal History Score	0.41	0.01	60.58	<0.001
White	-0.07	0.03	-2.61	0.009
Constant	-2.81	0.04	-69.71	<0.001
Indirect Effect	-0.10	0.003	-35.69	<0.001
Direct Effect	-0.02	0.006	-2.62	0.009
Total Effect	-0.11	0.006	-17.53	<0.001
Prop. Total Effect Of Race Mediated by Criminal History	0.85			
Ratio of Indirect to Direct Effect	5.85			
Ratio of Total to Direct Effect	6.85			

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

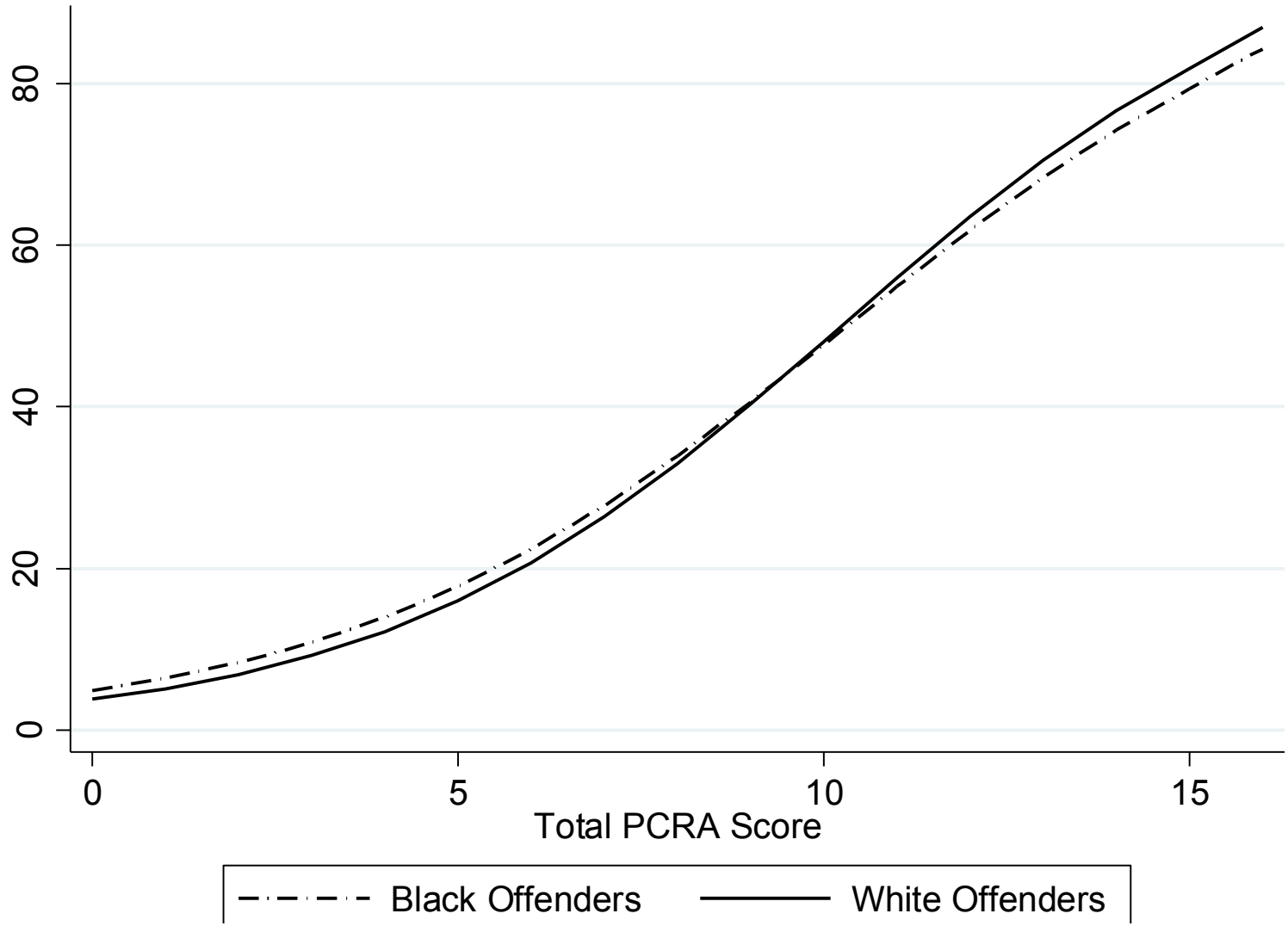


Figure 2. Rate of Re-Arrest for Any Crime and Percent Black by PCRA Score

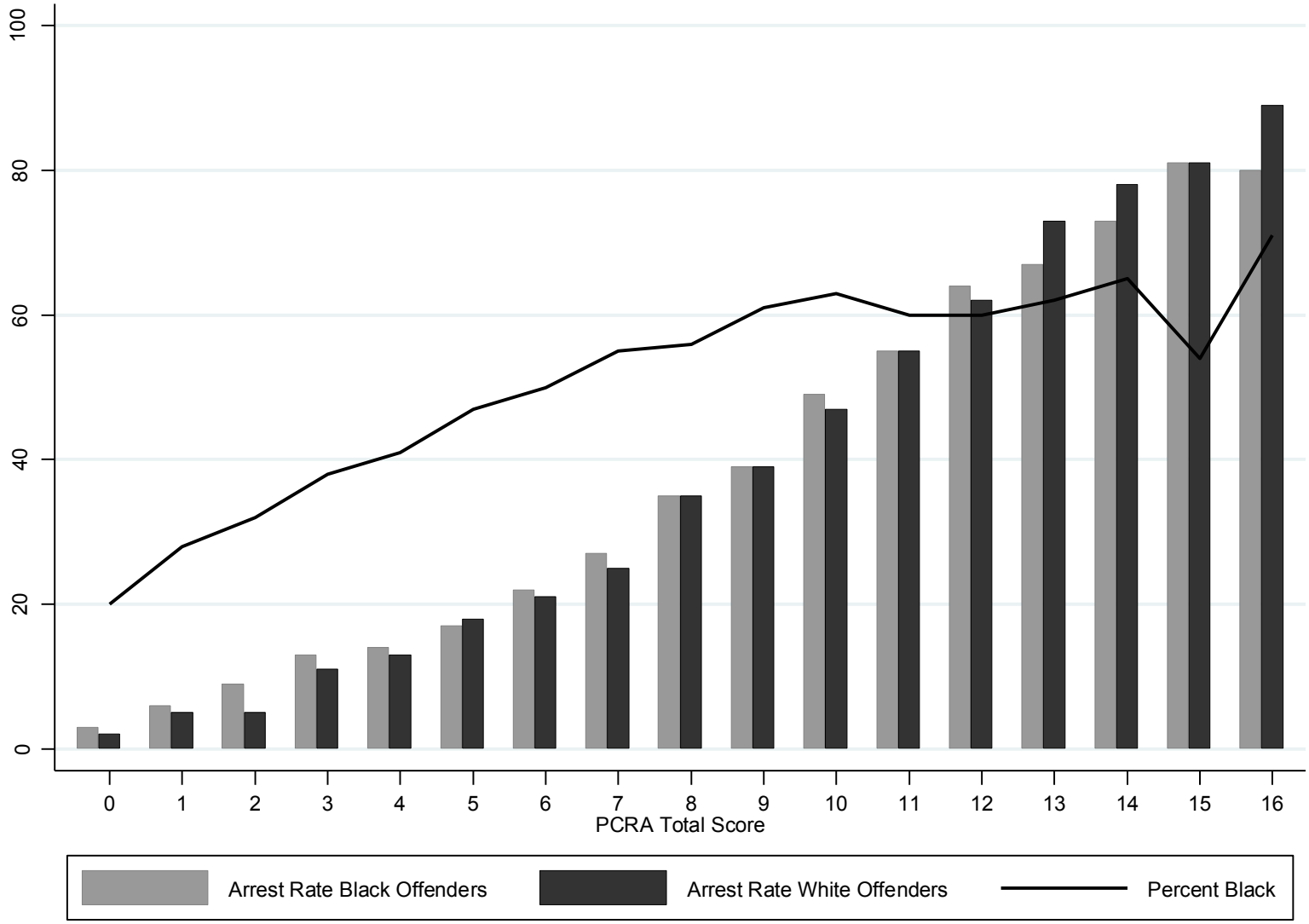
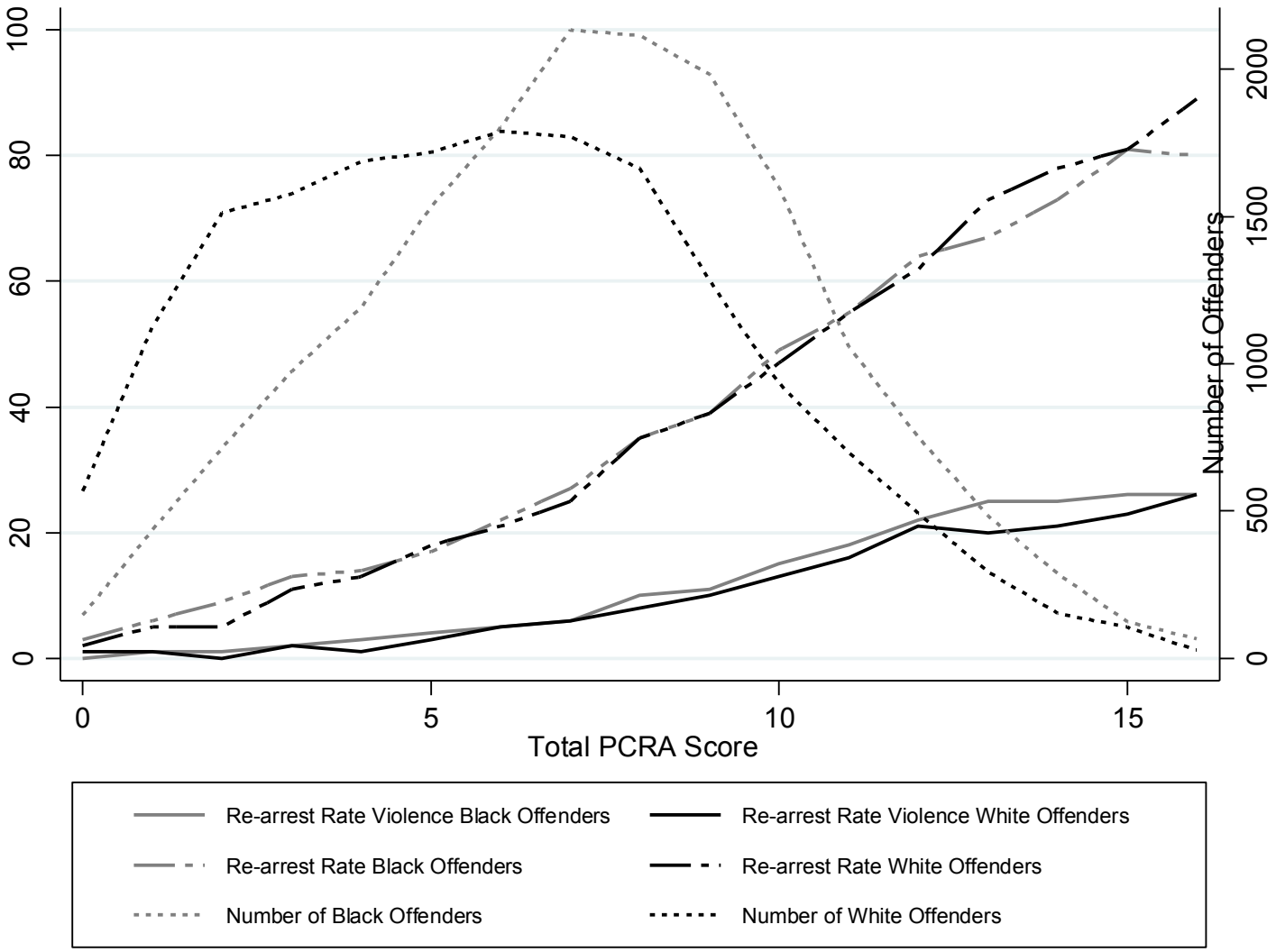


Figure 3. Rate of Re-Arrest for Any- and Violent- Crime and PCRA Distribution by Race



REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *The Standards for Educational and Psychological Testing.* Washington, DC: AERA Publications.

American Law Institute (2014). *Model Penal Code: Sentencing (Tentative Draft No. 3).* Philadelphia: American Law Institute.

Arnold, H. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior & Human Performance, 29* 143-174.

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. 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 .

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.

Ender, P.B. (2011). Binary_mediation: Command to compute indirect effect with binary mediator and/or binary response variable. UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/stata/ado/analysis/.

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. Accessed 10/10/15 at: 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.

Griggs v. Duke Power Co. (1971) 401 U.S. 424

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.

Herrenkohl, T. I., Maguin, E., Hill, K. G., Hawkins, J. D., Abbott, R. D., & Catalano, R. F. (2000). Developmental risk factors for youth violence. *Journal of Adolescent Health*, *26*(3), 176-186.

Hoge, R. D. (2002). Standardized instruments for assessing risk and need in youthful offenders. *Criminal Justice and Behavior, 29,* 380-396.

Holder, E. (2014). Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting. Available at: http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, *75*, 16-29.

Kim, H. S. (2010). *Prisoner classification re-visited: A further test of the Level of Service Inventory-Revised (LSI-R) intake assessment* (Doctoral dissertation, Indiana University of Pennsylvania).

Kleiman, M., Ostrom, B., & Cheesman, F. (2007). Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency, 53*, 106-132.

Kraemer, H.C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry* 158:848–856

Lawrence A. 2013. Trends in Sentencing and Corrections: State Legislation. Denver: National Conference of State Legislatures http://www.ncsl.org/Documents/CJ/TrendsInSentencingAndCorrections.pdf

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906-917.

Lowenkamp, C. T., & Bechtel, K. (2007). Predictive Validity of the LSI-R on a Sample of Offenders Drawn from the Records of the Iowa Department of Correction Data Management System. *Federal Probation*, *71*, 25-34.

Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA Revisited: Testing the Validity of the Federal Post Conviction Risk Assessment (PCRA). *Psychological Services, 12,* 149-157.

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The Federal Post Conviction Risk Assessment (PCRA): A construction and balidation 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. (in press). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology.*

Monahan, J., & Skeem, J. (2014). Risk redux: The resurgence of risk assessment in criminal sentencing. *Federal Sentencing Reporter, 26,* 158-166.

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.

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.

Pennsylvania Sentencing Commission (March, 2015). Special Report: Impact of Removing Demographic Factors. Retrieved 10/1015 from: http://pcs.la.psu.edu/publications-and-research/research-and-evaluation-reports/risk-assessment/phase-ii-reports/special-report-impact-of-removing-demographic-factors/view

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.

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

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.

Singh, J. P., & Fazel, S. (2010). Forensic Risk Assessment A Metareview. *Criminal Justice and Behavior*, *37*(9), 965-988.

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.

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.

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

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.

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* http://www.justice.gov/sites/default/files/criminal/legacy/2014/08/01/2014annual-letter-final-072814.pdf

Yang, M, Wong, S.C., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin* 136: 740-767

Endnotes

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

[ii] Prior to matching, the average age of Black and White offenders was 37.65 and 42.47 respectively ($t(44,717) = -48.48$; $p < 0.001$), and 88% and 82% of Black and White offenders, respectively, were male ($\chi^2(1) = 344.49$; $p < 0.001$). The correlation of race with age and sex in the unmatched sample would yield imprecise estimates of race effects (for Aims 2-3) and require more complex interaction terms in moderation models (for Aim 4)—which are important, but not the focus of this paper. By matching on age and sex, we focus more specifically on the relationship between risk and race. Note that all analyses were also completed with the full, non-matched sample—and yielded a similar pattern of results across aims (tables and figures available upon request).

[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] It should be noted that several additional and more complete moderation models were estimated. Those models controlled for interactions between race, sex, age and the PCRA and indicated that the simple interaction term between race and the PCRA was not statistically significant ($p \leq 0.01$) nor were the more complex interaction terms that included race. The model presented in Table 2 was selected because it is a simple model and had the best chance of highlighting potential differences, by race, in the functional form of the relationship between the PCRA and re-arrest.

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

[vi] A model using PCRA Total Score as the potential mediator yielded similar results—indicating that 90% of the effect of race on re-arrest was mediated by risk.

[vii] Theoretically, it is possible. Most validated risk assessment tools have predictive utilities that are essentially interchangeable (Yang, Wong & Coid, 200x). 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; Sackett et al., 2001)—particularly for short scales (see Pennsylvania Sentencing Commission, 2015).