RACIAL BIAS AND LSI-R ASSESSMENTS IN PROBATION SENTENCING AND OUTCOMES

EVAN M. LOWDER
Indiana University–Purdue University Indianapolis

MEGAN M. MORRISON
Tennessee State University

DARYL G. KRONER
Southern Illinois University–Carbondale

SARAH L. DESMARAIS
North Carolina State University

Risk assessments are now implemented in correctional settings across the United States as an evidence-based strategy to inform sentencing and supervision decisions. Despite growing research examining racial bias in the predictive validity of risk assessments, few studies have investigated racial bias in the context of judicial decision-making. We investigated the interactive contributions of race and Level of Service Inventory–Revised (LSI-R) risk assessments in predicting sentence length and probation outcomes in 11,792 Black and White probationers. Results showed White probationers at low-risk levels received longer sentences relative to Black probationers classified at the same risk levels. However, there were few differences at higher risk levels and no evidence of racial bias in the predictive accuracy of LSI-R assessments for other probation outcomes. Findings highlight the need for prospective and carefully controlled investigations into whether risk assessments improve the fairness and accuracy of sentencing and other risk management decisions.

Keywords: risk assessment; racial bias; probation; sentencing; predictive validity

INTRODUCTION

There are more than 4.5 million adults under correctional supervision at a given time in the United States, amounting to one in every 53 adults. Probationers account for more than 80% of adults under community supervision (Kaeble & Bonczar, 2016). Although the number of adults under correctional supervision has been relatively stable over the past decade, efforts to adopt evidence-based practices and other data-driven strategies to effectively and efficiently manage this population have expanded in recent years. In particular, risk assessment has emerged as a preferred evidence-based practice to inform sentencing and supervision decisions. Today, as many as 27 U.S. states incorporate risk assessment into data-driven

AUTHORS' NOTE: *We thank Bree Derrick with the Council of State Governments for her assistance in acquiring these data.* *Correspondence concerning this article should be addressed to Evan M. Lowder, School of Public and Environmental Affairs, Indiana University–Purdue University Indianapolis, 801 W. Michigan Street, BS 3025, Indianapolis, IN 46202; e-mail: elowder@iu.edu.*

CRIMINAL JUSTICE AND BEHAVIOR, 2019, Vol. 46, No. 2, February 2019, 210–233.
DOI: 10.1177/0093854818789977
Article reuse guidelines: sagepub.com/journals-permissions

justice initiatives to inform criminal justice decision-making (Lawrence, 2013). Risk assessments implemented in correctional contexts serve several purposes, including informing imprisonment, informing supervision and release decisions, and informing efforts to reduce risk (e.g., intensity of supervision; Monahan & Skeem, 2016).

RACIAL BIAS IN RISK ASSESSMENT

The growing use of risk assessments to inform correctional decision-making has not been met without controversy. Critics have voiced concerns over the potential for racial bias in the measurement of risk and subsequent application of risk estimates in correctional settings (Harcourt, 2015). To illustrate, in 2014, former U.S. Attorney General Eric Holder cautioned that risk assessments may undermine justice because they are often based on unchangeable, or static, factors that are biased toward minority offenders (U.S. Department of Justice, 2014). Accordingly, critics have raised concerns about the constitutionality of using risk assessments to inform sentencing and parole decisions (Starr, 2014).

Although race has not been measured overtly in risk assessments since the mid-1900s, the inclusion of criminal history and other socioeconomic factors in risk assessment has been argued to serve as a proxy for race due to a focus on largely immutable characteristics of an offender (Harcourt, 2015). However, some have criticized the labeling of such risk factors as "proxies," arguing that criminal history and other risk factors (isolated from race) are stronger predictors of recidivism relative to race alone (Monahan & Skeem, 2016). As such, risk factors would be better characterized as overlapping with race or mediating associations between race and recidivism (Monahan & Skeem, 2016). Regardless, the primary rationale for racial bias in risk assessments is centered on the inclusion of factors in the measurement of risk that—particularly for Black Americans—are relatively static across the lifespan (e.g., socioeconomic characteristics, criminal history) and disadvantage Black Americans relative to White Americans. Indeed, even the most commonly used risk assessments capitalize on these characteristics in predicting offending risk. A recent review of risk assessments employed in correctional settings found that all included measures of antisocial behavior (often measured by criminal history), most included current engagement in work or educational activities, and nearly half included measures of housing (Desmarais, Johnson, & Singh, 2016).

Prior research has provided evidence of systematic racial disparities in socioeconomic factors and criminal involvement. For example, relative to their White counterparts, Black Americans have lower levels of wealth accumulation (Proctor, Semega, & Kollar, 2016; U.S. Census Bureau, 2017b; Vornovitsky, Gottschalck, & Smith, 2011), lower rates of homeownership (Callis & Kresin, 2016), and are less likely to graduate high school and complete a 4-year degree (U.S. Census Bureau, 2017a). In correctional contexts, Black Americans are more likely to be stopped by police (Gelman, Fagan, & Kiss, 2007), incarcerated (Abrams, Bertrand, & Mullainathan, 2012; Bales & Piquero, 2012; Kutateladze, Andiloro, Johnson, & Spohn, 2014), and charged with a crime (Wu, 2016) relative to White Americans. Furthermore, after initial detention, Black Americans are less likely to be released on their own recognizance and, when they are released, more likely to receive higher bond amounts (Wooldredge, 2012). At conviction, racial minorities receive harsher criminal sentences (Sweeney & Haney, 1992) relative to their nonminority counterparts, particularly for person crimes (Kutateladze et al., 2014). Together, these trends contribute

to more substantial criminal histories for racial minorities. In the context of risk assessment, these trends suggest that Black offenders will receive risk assessment scores that are much higher than—and disproportionate to—their actual risk of offending, due to relative disadvantage Black Americans have on these factors relative to White Americans. In this way, risk assessments used to inform correctional decisions may both capitalize on and exacerbate social inequalities (van Eijk, 2016).

THE LEVEL OF SERVICE INVENTORY–REVISED (LSI-R)

The LSI-R (Andrews & Bonta, 1995, 2001) is an actuarial instrument focused on the measurement of both dynamic (i.e., changeable) and static risk factors and designed to predict risk of general offending (Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990; Andrews, Bonta, & Wormith, 2010). Developed from the Risk-Need-Responsivity model of effective offender rehabilitation (Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990), the LSI-R is now used by more than 900 correctional agencies in practice (Lowenkamp, Lovins, & Latessa, 2009), making it one of the most widely used risk assessments in correctional settings. The instrument includes 54 items covering 10 risk domains: Criminal History, Education/Employment, Financial, Family/Marital, Accommodation, Leisure/Recreation, Companions, Alcohol/Drug Problems, Emotional/Personal, and Attitudes/Orientation. Resulting total scores are categorized into five risk bins: Low, Low-Moderate, Moderate, Moderate-High, and High.

LSI-R assessments conducted in correctional settings have shown some evidence of inter-rater agreement (Andrews et al., 2010; Lowder, Desmarais, Rade, Johnson, & Van Dorn, 2017; Lowenkamp, Holsinger, Brusman-Lovins, & Latessa, 2004), though the reliability of assessments has been found to vary among racial and ethnic minorities (Schlager & Simourd, 2007). A substantial body of research has focused on establishing evidence for the predictive validity of LSI-R assessments with respect to general recidivism, with some success (Olver, Stockdale, & Wormith, 2014; Singh, Grann, & Fazel, 2011; Vose, Cullen, & Smith, 2008). However, in comparative investigations, LSI-R assessments have been found to produce weaker predictive validity estimates relative to risk assessments produced by other instruments (Desmarais et al., 2016; Singh, Desmarais, & Van Dorn, 2013).

Given growing concerns about the potential for racial bias in assessments of recidivism risk, several studies have investigated differences in the predictive validity of LSI-R assessments as a function of race. In one study of 1,910 White and 672 Black prison inmates, LSI-R assessments produced weaker predictive validity estimates for institutional misconduct risk over a 2-year period for non-White prison inmates relative to White inmates (Chenane, Brennan, Steiner, & Ellison, 2015). Another investigation of 5,647 Black and 2,455 White offenders released from prison found weaker estimates of re-arrest and parole revocation risk over a 1-year period for Black offenders relative to White offenders (Ostermann & Salerno, 2016). A third study of 696 African American and 133 Caucasian male offenders in two treatment facilities in New Jersey found that LSI-R total scores overclassified African American offenders at higher risk of re-arrest relative to Caucasian offenders, who were more likely to be underclassified (i.e., receive false negative results). However, overall predictive validity estimates were slightly stronger for African American offenders relative to Caucasian offenders (Fass, Heilbrun, DeMatteo, & Fretz, 2008). A final investigation in a sample of 43 African American and 52 Caucasian mental health

diversion program participants found that African American offenders assessed at high-risk had fewer days incarcerated over a 3-month period—but not 6- to 18-month follow-up periods—relative to White offenders assessed at the same risk level (Lowder et al., 2017).

Furthermore, the choice of both LSI-R cutoff scores and outcome measures employed in correctional settings has been found to affect classification errors. In one study of 279 Black and 177 White offenders in a residential work facility, an LSI-R cutoff score of 16 was found to produce substantial overclassification errors for Black offenders relative to their rate of program completion and relative to an alternative cutoff score of 25 (Whiteacre, 2006). Indeed, the selection of cutoff scores and associated risk categories (e.g., a 3-category, 4-category, or 5-category approach) across risk assessment instruments implemented in correctional settings has been criticized as potentially arbitrary and a hindrance to efforts to adopt more standardized approaches to the classification and interpretation of risk (Hanson et al., 2017). In 2014, the Council of State Governments Justice Center convened a series of meeting of researchers, policymakers, and correctional practitioners to craft a common language for the interpretation of risk across risk assessment instruments. The resulting product was a five-level approach to the classification of risk based on standardized predicted probabilities of offending at each level that could be adopted universally with any risk assessment instrument (Hanson et al., 2017). However, whether a standardized approach to risk assessment classification could mitigate racial bias in risk assessment remains to be seen.

FAIRNESS IN RISK ASSESSMENT

Given heightened concerns about bias in risk assessment and in light of highly publicized reports (Angwin, Larson, Mattu, & Kirchner, 2016), recent efforts have focused on providing conceptual and operational definitions of fairness in an effort to refine and generally improve our shared understanding of what is meant by "bias" in risk assessment. To illustrate, Kleinberg, Mullainathan, and Raghavan (2016) proposed three properties of fair risk assessments. First, risk assessments should be calibrated within groups such that scores (and associated risk classifications) have similar meanings for each group (i.e., associated with a similar increase in risk). Second, risk assessments should be balanced for the negative class such that among those who do not offend, scores should be similar between groups. Third, risk assessments should be balanced for the positive class such that among those who do offend, scores are similar between groups. These properties lend themselves to important conclusions, including that when base rates (e.g., rates of offending) differ between groups, it is impossible to satisfy all three definitions of fairness.

Other authors have expanded on these properties by providing specific definitions of algorithmic fairness, though still concluding that total fairness is impossible to achieve in practice (Berk, Heidari, Jabbari, Kearns, & Roth, 2017). Furthermore, deciding on which standards of fairness to prioritize comes with necessary tradeoffs—that is, satisfying fairness at the cost of reducing public safety or, conversely, improving public safety at the cost of compromising fairness (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). These tradeoffs are inherent to the use of risk assessments in practice and are issues that must be addressed by criminal justice stakeholders, not researchers (Berk et al., 2017).

Perhaps more relevant to the use of risk assessments in criminal justice practice—where risk assessment information is primarily used as a mitigating consideration within the

bounds of the law rather than a clear indicator of a "positive" or "negative" test—are fairness definitions articulated by Skeem and Lowenkamp (2016). Specifically, these authors differentiate between predictive bias, whereby an instrument produces different levels of predictive accuracy across racial groups, and disparate impact, whereby an instrument that produces higher scores for one group versus another leads to unfair application of risk assessment scores in practice or even the perception of unfair decision-making. These definitions are consistent with the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing [U.S.], 2014), which argue that a test should be free of predictive bias. That is, a fair test should produce assessments that are similarly accurate in predicting an outcome across groups. In the context of sentencing, we argue further that risk assessment results should be applied consistently across groups, regardless of their predictive accuracy. This is similar to Skeem and Lowenkamp's (2016) conceptualization of disparate impact (consistent with Standards 3.6 and 3.20), but differs in that it focuses specifically on the disparate application of risk assessment findings to criminal justice decision-making rather than the perception of disparate application.

Consistent with these definitions—and for the purposes of the present investigation—operationally we define a lack of fairness as evidence of a statistically significant and practically meaningful moderation effect of race by risk assessment on sentencing and community supervision outcomes. It is not sufficient to demonstrate that Black probationers receive higher scores or risk classifications or have higher rates of recidivism relative to White probationers. These issues are indeed problematic and introduce other sources of bias into the interpretation and application of risk assessment results, which others have explored (Berk et al., 2017; Skeem & Lowenkamp, 2016). However, consistent with concerns in the legal community regarding racial bias in risk assessment (e.g., Harcourt, 2015; U.S. Department of Justice, 2014; Starr, 2014), assessment bias will be determined by overclassification of Black probationers at higher risk levels and underclassification at lower risk levels, relative to their rate of offending and relative to White probationers classified at the same risk level. Stated differently, risk assessments should provide similar ability to discriminate between risk classifications for different racial groups, regardless of the base rate of offending in each group.

STUDY PURPOSE

Despite a handful of studies investigating racial bias in the predictive validity of LSI-R assessments, to our knowledge there have been few investigations of racial bias in the context of correctional decision-making, particularly in the United States. Yet, risk assessments have potential for racial bias both in their ability to predict likelihood of an outcome similarly across racial groups and in their use to inform correctional decisions regarding sentencing, supervision, and treatment. The application of risk assessment instruments in the context of sentencing, in particular, may bias racial minorities toward longer sentences and greater criminal justice involvement. However, to date, few studies have evaluated the use of risk assessments in this context. Moreover, there has been limited investigation of racial bias in risk assessments in probationers, specifically, who make up a considerable portion of adults under correctional supervision.

To address these limitations, we examined several empirical trends in a state-wide sample of offenders sentenced to probation. Primarily, we investigated predictive associations between LSI-R assessments and sentencing and community outcomes. In accordance with the definition of disparate impact, we first investigated the independent and interactive contributions of race and LSI-R assessments to the prediction of sentence length. Second, consistent with the definition of predictive bias, we examined between-group differences in the validity of LSI-R assessments in predicting failure to complete probation and acquiring a new charge on probation. Third, and additionally consistent with the definition of disparate impact, we evaluated how application of a five-level risk assessment approach based on alternative cutoff scores affected risk classification for Black and White probationers.

METHOD

SAMPLE

The sample comprised 11,792 probationers in Kansas who were sentenced to probation only and who were primarily White (n = 8,811, 74.7%) versus Black (n = 2,981, 25.3%). Most were male (n = 9,276, 80.1%) and non-Hispanic (n = 10,404, 88.2%). Participants were an average age of 32.37 years (SD = 11.26 years, range = 15-84 years) at the time of assessment. The majority of probationers were not currently married (77.0%, n = 9,076). Only 28.4% (n = 3,349) had graduated high school or received a General Education Diploma (GED), 59.4% (n = 7,006) had some high school education, and 3.8% (n = 438) had less than a high school education. Index offenses were primarily for nonperson (64.3%, n = 7,579) crimes. Participants had an average of 1.16 (SD = 0.67, range = 1-21) charges associated with the index offense, corresponding to an average 5.99 (SD = 2.29, range = 1-10) severity level.

PROCEDURES

We obtained court probation sentencing and LSI-R assessment records from the Kansas Department of Corrections for probationers supervised from 2003 to 2015. Similar to other jurisdictions that require the use of LSI-R assessments in presentence investigations to inform sentencing decisions (e.g., Casey et al., 2013; Monahan & Skeem, 2016), Kansas state policy mandates the LSI-R as a component of the presentence investigation for all felony cases and some misdemeanor cases to inform probation sentencing (Rule 110B of KSA 75-5291[a][2]). Assessment results determine whether someone is sentenced to community corrections versus court-based supervision (KSA-75-5291) and are required to be accessible to the judge as a component of the presentence investigation (KSA-21-6813). Thus, assessment information is available as a mitigating consideration for sentencing within the range of a presumptive probation sentence, consistent with Kansas sentencing grids.

Presentencing assessments were matched to sentencing records based on sentencing dates occurring within 45 days of assessment, consistent with state policy (KSA 75-5291[a][2]). Because date of assessment indicated the date that the assessment was entered into the administrative database (i.e., not the date that the assessment was conducted), we included any assessment that was labeled presentencing and could be matched to a sentencing date within 45 days. Overall, 28,890 presentencing assessments were matched to sentencing dates. However, consistent with the use of LSI-R assessments to inform probation

sentencing decisions, we refined our sample to defendants who were sentenced to probation only, whose charges fell under the scope of Kansas sentencing grids, who identified as Black or White, and who had a closed probation case. Our final sample consisted of 11,792 primarily felony-level defendants who received a presentencing LSI-R and were subsequently sentenced for community supervision.

MEASURES

Outcome Variables

Our outcomes included sentence length, any probation failure, and any new charge. These outcomes were selected to align with the purpose of the presentencing LSI-R (i.e., used in the context of judicial decision-making to inform the court-mandated community supervision sentence). Sentence length (months) measured the length of probation as determined at sentencing. For probation outcomes, we investigated any failure (yes, no), indicating whether a probation sentence was terminated for any reason or successfully completed. Reasons for termination included a closed case, a remanded case, or revocation. In addition, new charge (yes, no) indicated receipt of a new felony or misdemeanor charge while on probation. Primarily, new charge was coded from court probation records, indicating whether the case was revoked for a new misdemeanor or felony charge. However, we additionally consulted court conviction and sentencing records, including any offense that occurred during the probationary period and resulted in a criminal charge and subsequent conviction.

Although arrest or jail booking data were not available, our selection of recidivism measures is consistent with the purpose of the presentencing assessment. In this context, the most relevant outcomes include the likelihood that a person will fail to complete probation successfully and the likelihood that a person will recidivate to court with a new charge. Although arrest is commonly employed as a measure of general offending in risk assessment research (Desmarais et al., 2016), it is a more generic measure of general offending that is less specific to the probation context. For example, an arrest could be associated with a new charge or another type of probation violation. Thus, our choice of outcomes reflects the use of the LSI-R for a specific purpose (i.e., the probation presentencing process) rather an attempt to establish the predictive accuracy of LSI-R assessments for general offending outcomes broadly.

LSI-R

The LSI-R is a risk and needs assessment designed to predict risk of general offending and technical violations in adult offenders (Andrews & Bonta, 2001). The instrument includes 54 items measuring a range of static and dynamic risk factors. Items are scored indicating absence (0) or presence (1) of a risk factor, the scores of which are totaled to create an LSI-R total score ranging from 0 to 54. Total scores classify offenders into five risk bins: Low (0-13), Low-Moderate (14-23), Moderate (24-33), Moderate-High (34-40), and High (41-54). In this study, we examined both LSI-R total scores and risk bins or classifications. Because we received LSI-R total scores for probationers, not item-level data, we could not compute internal consistency estimates. However, Kansas state policy requires comprehensive training of court officers who complete LSI-R assessments, including 6 hr

of training following initial training and refresher training if an LSI-R assessment has not been completed in 6 months. Furthermore, LSI-R assessments have been found to have good levels of inter-rater reliability when conducted by criminal justice professionals in other jurisdictions (Simourd, 2006).

Race

Race was a primary variable of interest and was measured dichotomously (White, Black), separate from ethnicity. Participants who identified as a race other than Black or White were excluded from analysis.

Covariates

Several variables were included as covariates in all multivariate models. Offense severity (1-10) was based on Kansas sentencing grids, with higher numbers indicating less severe offenses. Counts (count) indicated the number of charges associated with the index offense. Type of index offense indicated whether the index offense was for a person or nonperson crime. Age (continuous) and sex (male, female) were additionally included as participant characteristics. Time at risk (days from probation sentencing to termination date) was used as a covariate in all models predicting probation failure and new charge.

ANALYSES

Prior to addressing each aim, we conducted descriptive statistics on all study variables to assess missingness, response distributions, and assumptions of normality. There was a small and negligible proportion of missing data across study variables (0.16%; Schafer, 1999), primarily due to missing demographic information. This resulted in an analytic sample of 11,585 participants in multivariable models. In addition, we found some evidence of skew for sentence length (skewness = 1.51). However, general linear approaches can tolerate skew values more than 2 (Tabachnick & Fidell, 2013). As a result, we modeled sentence length using a general linear approach. Both community supervision outcomes were measured dichotomously and showed good variability in response distributions.

MULTICOLLINEARITY TESTING

Prior to the main analyses, we assessed for evidence of potential multicollinearity between predictors using bivariate comparisons (i.e., Pearson correlations, chi-square tests of independence, and one-way ANOVAs). Pearson correlation results showed associations below published thresholds for multicollinearity (r = .90; Tabachnick & Fidell, 2013). The strongest correlation was observed between time on probation and crime type, r (11790) = .24, $p < .001$. Results of chi-square tests showed the strongest association between sex and crime type, $\chi^2(1) = 207.07$, $p < .001$, $\Phi = .13$; however, this association was small in magnitude (Cohen, 1988). Similarly small associations were observed for results of one-way ANOVAs. The strongest association was observed between time on probation and LSI-R risk classification, $F(4, 11791) = 139.32$, $\eta^2 = 0.04$, though this association was only small-to-medium in magnitude (Cohen, 1988). Overall, these findings provided little evidence of the potential for multicollinearity in multivariable models. As such, all predictors were included in subsequent models.

MULTILEVEL MODELS

To address the primary study aims, we employed multilevel modeling (MLM) to account for the nested structure of study data (Raudenbush & Bryk, 2002). All analyses were conducted using the PROC MIXED command and GLIMMIX macro (Guo & Zhao, 2000) in SAS 9.4. MLM has an advantage over other statistical approaches in that it does not require a similar number of observations in each larger grouping and can be used even when observations are missing, making it a suitable analytic strategy for administrative data where observations are nested within geographic regions. Importantly, MLM allows for variability in outcomes to be isolated at two or more levels of analysis and predictors used to explain variability at each level. In the present study, probationer-level observations (Level 1) were nested within counties at (Level 2). Although the purpose of this study was to explain between-probationer (i.e., Level 1) variability in probation sentencing and supervision outcomes, parsing out variability at Level 2 (i.e., between-counties) increased the validity of our Level 1 inferences by controlling for jurisdictional differences.

Prior to conducting multivariable analyses, preliminary analyses must be conducted to establish significant variability at each level of analysis to justify the overall MLM approach and inclusion of predictors (Nezlek, 2001; Raudenbush & Bryk, 2002). Thus, we first conducted preliminary analyses using unconditional null models with no predictors to establish significant Level 1 and Level 2 variability for each of our three outcomes.

Subsequently, to test for evidence of an interaction effect, we employed hierarchical regression analysis using multilevel models. Hierarchical regression analysis (distinct from hierarchical linear modeling, another term for MLM) refers to the testing of regression models in sequence where model fit is evaluated following the addition of new model terms. In the present study context, this refers to testing for improvement in model fit from a main-effects only model to a model with main effects and an additional interaction term. Accordingly, in Block 1, we added all predictors as main effects. In Block 2, a race by LSI-R term was added to examine the interactive effect of race and LSI-R on outcomes, controlling for predictors. We replicated this model using both LSI-R risk classifications and LSI-R total scores, separately, as well as for each dependent variable. All models controlled for covariates as previously defined. Furthermore, prior to analysis, all continuous predictors were group-mean centered (i.e., at county mean levels). In this approach, the intercept for multilevel models becomes the expected outcome value when each group is at its mean level. Thus, for this study, centering reduced the likelihood that between-county differences would contribute to parameter estimates, thereby reducing bias in Level 1 predictors and nonessential multicollinearity in moderation analyses. Models of sentence length were conducted using PROC MIXED for a general linear model whereas models of probation failure and new charge were conducted using the GLIMMIX macro for dichotomous outcomes.

For all hierarchical regression models, we assessed model fit by examining change in −2 log likelihood statistics and comparing results to a chi-square distribution to evaluate statistical significance. To adjust for the possibility of Type I error across multiple tests, we used $\alpha = .01$ as our level of statistical significance, which produced equivalent results as applying a Holm-Bonferonni correction for multiple comparisons (Holm, 1979). For all effects, we present unstandardized regression coefficients (*B*) and standard errors (*SE*). Although standardized regression coefficients are used frequently in single-level regression models, they are more challenging to compute and require careful interpretation in multilevel models (Heck & Thomas, 2008). For this reason, and to facilitate cross-study comparisons (Hox,

1994; Hox, Moerbeek, & van der Schoot, 2010), we report unstandardized coefficients only. For dichotomous outcomes, we additionally present odds ratios (ORs) and associated 99% confidence intervals, consistent with our level of significance ($\alpha = .01$). Where significant, we present group means adjusted for mean levels of continuous predictors, decomposed interactions (i.e., estimated slopes and contrasts), and associated effect size estimates (i.e., standardized mean difference, *d*; Cohen, 1988). Where significant, interactions involving continuous predictors were probed at +/– 1 SD above and below mean values.

POST HOC COMPARISONS USING MODIFIED RISK CLASSIFICATIONS

In further investigation of disparate impact, we conducted a series of comparisons to determine whether a modified risk classification system would change the frequency distribution of risk levels overall and by race. Cutoff values for modified risk classifications were calculated consistent with the five-level approach (Hanson et al., 2017) and using the entire sample of Kansas probationers to establish accurate base rates of offending. The minimum LSI-R score (1) and maximum LSI-R score (54) established the lower and upper bounds for Levels I and V, respectively. To establish upper bound cutoff values for Levels I and IV, 5% and 85% offending probabilities were used, respectively (Babschishin, Kroner, & Hanson, 2017). To establish the upper bound cutoff values for Levels II and III, odds of .70 and 1.43 were used, respectively. These values represented the average treatment effect associated with criminal justice interventions, or $r = .10$, $d = .20$, or an OR of .70/1.43 (Andrews, Zinger, et al., 1990). We used logistic regression to generate predicted probabilities of offending over a 2-year follow-up for each individual LSI-R score to inform cutoff values for Levels I and V (i.e., where the predicted probability reached 5% and 85%, respectively; Helmus & Hanson, 2011). However, no predicted probability of offending reached 85%; as a result, no cases were assigned to Level V. From this model, we additionally computed odds of offending at each LSI-R score to identify upper bound cutoff values of Levels II and III. The resulting cutoff values produced four risk levels (Level I: 1-5; Level II: 5-13; Level III: 14-34; Level IV: 35-50), which were subsequently applied to the current sample. We then examined change in risk classification from the LSI-R original risk classification to the five-level risk system. Frequencies and chi-square tests of independence were conducted to examine change in risk classifications overall and as a function of race, respectively.

RESULTS

DESCRIPTIVE

LSI-R Assessments

Participants were primarily classified as Low-Moderate ($n = 4{,}270$, 36.2%) or Moderate ($n = 4{,}577$, 38.8%) risk on the LSI-R, with fewer participants classified at Low ($n = 1{,}402$, 11.9%), Moderate-High ($n = 1{,}319$, 11.2%), or High ($n = 224$, 1.9%) risk levels. Consistent with risk bin classification, LSI-R total scores averaged 23.88 ($SD = 8.34$, range = 1-50), corresponding to a Moderate risk classification.

Outcomes

Participants were sentenced to an average 19.74 months on probation ($SD = 8.06$ months, range = 12-60 months), but spent an average 15.76 months ($SD = 9.10$ months, range =

0-127 months) serving time on probation. Slightly less than half of participants were terminated from probation without successful completion (n = 4,759, 40.4%). A substantial portion of participants received a new charge from an offense committed during time on probation (n = 1,852, 15.7%).

BIVARIATE COMPARISONS

Consistent with our investigation of disparate impact, we examined bivariate comparisons between race and LSI-R risk classifications, LSI-R total scores, and outcome variables. Bivariate comparisons showed small but significant effects of race on LSI-R risk bin classification, $\chi^2(4) = 29.04$, $p < .001$, $\Phi = .05$, or LSI-R total scores, $t(5, 486.84) = 2.30$, $p = .021$, $d = 0.05$. Specifically, White participants were slightly more likely to be classified at Low (n = 1,117, 12.7%) and High (n = 183, 2.1%) risk relative to Black probationers (n = 41, 1.4% and n = 92, 1.4%, respectively). In contrast, Black probationers were more likely to be classified at Moderate risk (n = 1,219, 40.9%) relative to White probationers (n = 3,358, 38.1%). There were no differences between White and Black probationers in likelihood of classification at Low-Moderate (n = 3,168, 36.0% and n = 1,102, 37.0%, respectively) and Moderate-High (n = 985, 11.2% and n = 334, 11.2%) risk classifications. Black probationers additionally had a slightly higher average LSI-R total score (M = 24.17, SD = 7.88) relative to White probationers (M = 23.78, SD = 8.48). Regarding differences in outcome variables, there were no differences in average sentence length between White (M = 19.75, SD = 8.12) and Black (M = 19.71, SD = 7.87) probationers. Black probationers were more likely to have a sentence terminated for any reason (n = 1,474, 49.4%) compared with White probationers (n = 3,285, 37.3%), $\chi^2(1) = 136.91$, $p < .001$, $\Phi = .11$. However, Black probationers were only slightly more likely to have a new charge during probation (n = 579, 19.4%) compared with White probationers (n = 1,273, 14.4%), $\chi^2(1) = 41.64$, $p < .001$, $\Phi = .06$.

UNCONDITIONAL MODELS

Unconditional model results for sentence length showed significant variability at both Level 1 (B = 62.20, SE = 0.77, $p < .001$) and Level 2 (B = 1.52, SE = 0.41, $p < .001$). However, variability in sentence length existed primarily at Level 1 (i.e., between-probationers, 97.6%) versus Level 2 (i.e., between-counties, 2.4%). Although percent variability is not computed for nonlinear outcomes (Raudenbush & Bryk, 2002), unconditional models of probation failure and new charge showed significant variability at both Level 1 (B = 0.99, SE = 0.01, $p < .001$ and B = 0.98, SE = 0.01, $p < .001$) and Level 2 (B = 0.16, SE = 0.04, $p < .001$ and B = 0.12, SE = 0.04, $p < .001$). Furthermore, for both outcomes, Level 1 (i.e., between-probationer) variability was greater than Level 2 (i.e., between-county) variability. Consistent with these results, subsequent analyses focused on explaining between-probationer variability in outcomes, adjusting for variability attributable to between-county differences.

INTERACTIVE EFFECTS OF RACE AND RISK ASSESSMENT

Results of hierarchical regression analyses for LSI-R risk bins and total scores are presented in Tables 1 and 2, respectively.

TABLE 1: Summary of Hierarchical Regression Results by Outcome—LSI-R Risk Classifications

Variable	Sentence length (N = 11,585)	Any failure (N = 11,585)			New charge (N = 11,585)		
	B (SE)	*B (SE)*	OR	99% CI	*B (SE)*	OR	99% CI
Block 1							
Fixed-effects estimates							
Crime type (nonperson)	8.78*** (0.13)	0.45*** (0.07)	1.57	[1.31, 1.88]	0.14 (0.06)	1.16	[1.00, 1.34]
Severity	−1.20*** (0.03)	0.09*** (0.01)	1.09	[1.06, 1.14]	0.09*** (0.01)	1.09	[1.06, 1.12]
Counts	0.93*** (0.09)	0.13** (0.05)	1.13	[1.00, 1.28]	0.05 (0.03)	1.05	[0.96, 1.14]
Age	−0.04*** (<0.01)	−0.03*** (<0.01)	0.97	[0.96, 0.97]	−0.03*** (<0.01)	0.07	[0.96, 0.97]
Sex (male)	−0.11 (0.15)	−0.38*** (0.08)	0.68	[0.56, 0.84]	−0.41*** (0.07)	0.66	[0.55, 0.80]
Race (White)	−0.25 (0.15)	0.47*** (0.07)	1.59	[1.32, 1.93]	0.28*** (0.06)	1.32	[1.13, 1.54]
LSI-R (Low)							
Low-Moderate	−1.80*** (0.21)	0.92*** (0.13)	2.50	[1.81, 3.48]	0.60*** (0.12)	1.83	[1.34, 2.49]
Moderate	−1.93*** (0.21)	1.71*** (0.13)	5.52	[3.99, 7.65]	1.09*** (0.12)	2.98	[2.20, 4.04]
Moderate-High	−2.38*** (0.26)	2.38*** (0.15)	10.81	[7.35, 15.88]	1.39*** (4.03)	4.03	[2.87, 5.64]
High	−2.41*** (0.48)	2.77*** (0.26)	15.97	[8.09, 31.55]	1.25*** (0.20)	3.49	[2.08, 5.87]
Probation length		−0.07*** (<0.01)	0.93	[0.92, 0.94]	−0.01** (<0.01)	0.99	[0.98, 1.00]
Random-effects estimates							
Between-county residual variability	1.63*** (0.41)	0.12*** (0.04)			0.12** (0.04)		
Between-probationer residual variability	42.21*** (0.56)	1.99*** (0.03)			0.98*** (0.01)		
Block 2							
Fixed-effects estimates							
LSI-R (Low) × Race (White)							
Low-Moderate	1.78*** (0.50)	−0.09 (0.28)	0.91	[0.44, 1.87]	0.09 (0.26)	1.10	[0.56, 2.17]
Moderate	1.29** (0.50)	−0.31 (0.28)	0.73	[0.36, 1.50]	−0.18 (0.26)	0.83	[0.43, 1.62]
Moderate-High	2.01*** (0.61)	−0.18 (0.34)	0.84	[0.35, 1.99]	−0.14 (0.28)	0.87	[0.42, 1.81]
High	1.97 (1.21)	−0.10 (0.70)	0.91	[0.15, 5.51]	−0.81 (0.52)	0.44	[0.12, 1.70]
Random-effects estimates							
Between-county residual variability	1.64*** (0.41)	0.12*** (0.04)			0.13** (0.04)		
Between-probationer residual variability	42.16*** (0.56)	2.04*** (0.03)			0.98*** (0.01)		
Δ-2LL	15.30** (4)	3.90 (4)			7.87 (4)		

Note. Regression coefficients are unstandardized. For categorical variables, reference group indicated in parentheses. For severity, higher scores indicate lower severity. Δ-2LL reflects improvement in model fit upon addition of the interaction term(s) in Block 2. All model terms from Block 1 were included in Block 2; however, only unique terms are shown. LSI-R = Level of Service Inventory–Revised; CI = confidence interval for odds ratios; OR = odds ratio.
**p < .01. ***p < .001.

TABLE 2: Summary of Hierarchical Regression Results by Outcome–LSI-R Total Scores

Variable	Sentence length (N = 11,585)	Any failure (N = 11,585)			New charge (N = 11,585)		
	B (SE)	*B (SE)*	OR	99% CI	*B (SE)*	OR	99% CI
Block 1							
Fixed-effects estimates							
Crime type (nonperson)	8.81*** (0.13)	0.45*** (0.07)	1.57	[1.32, 1.86]	0.14 (0.06)	1.15	[0.99, 1.33]
Severity	−1.20*** (0.03)	0.09*** (0.01)	1.09	[1.06, 1.13]	0.09*** (0.01)	1.09	[1.06, 1.12]
Counts	0.93*** (0.09)	0.13** (0.04)	1.14	[1.01, 1.28]	0.05 (0.03)	1.05	[0.96, 1.14]
Age	−0.04*** (<0.01)	−0.03*** (<0.01)	0.97	[0.96, 0.97]	−0.03*** (<0.01)	0.97	[0.96, 0.97]
Sex (male)	−0.07 (0.15)	−0.42*** (0.07)	0.66	[0.54, 0.80]	−0.43*** (0.65)	0.65	[0.54, 0.78]
Race (White)	−0.27 (0.15)	0.46*** (0.07)	1.59	[1.33, 1.90]	0.28*** (0.06)	1.33	[1.14, 1.55]
LSI-R total score	−0.07*** (0.01)	0.09*** (<0.01)	1.10	[1.08, 1.11]	0.05*** (<0.01)	1.05	[1.04, 1.06]
Probation length		−0.07*** (<0.01)	0.93	[0.93, 0.94]	−0.01 (<0.01)	0.99	[0.98, 1.00]
Random-effects estimates							
Between-county residual variability	1.49*** (0.38)	0.11*** (0.04)			0.13** (0.04)		
Between-probationer residual variability	42.33*** (0.56)	1.76*** (0.02)			0.97*** (0.01)		
Block 2							
Fixed-effects estimates							
LSI-R total score × Race (White)	0.05** (0.02)	−0.01 (0.01)	0.99	[0.97, 1.01]	−0.01 (0.01)	0.99	[0.97, 1.01]
Random-effects estimates							
Between-county residual variability	1.48*** (0.38)	0.11*** (0.04)			0.13** (0.04)		
Between-probationer residual variability	42.30*** (0.56)	1.79*** (0.02)			0.97*** (0.01)		
Δ-2LL	7.20** (1)	1.87 (1)			2.16 (1)		

Note. Regression coefficients are unstandardized. For categorical variables, reference group indicated in parentheses. For severity, higher scores indicate lower severity. Δ-2LL reflects improvement in model fit upon addition of the interaction term(s) in Block 2. All model terms from Block 1 were included in Block 2; however, only unique terms are shown. LSI-R = Level of Service Inventory–Revised; CI = confidence interval for odds ratios; OR = odds ratio.
**p < .01. ***p < .001.

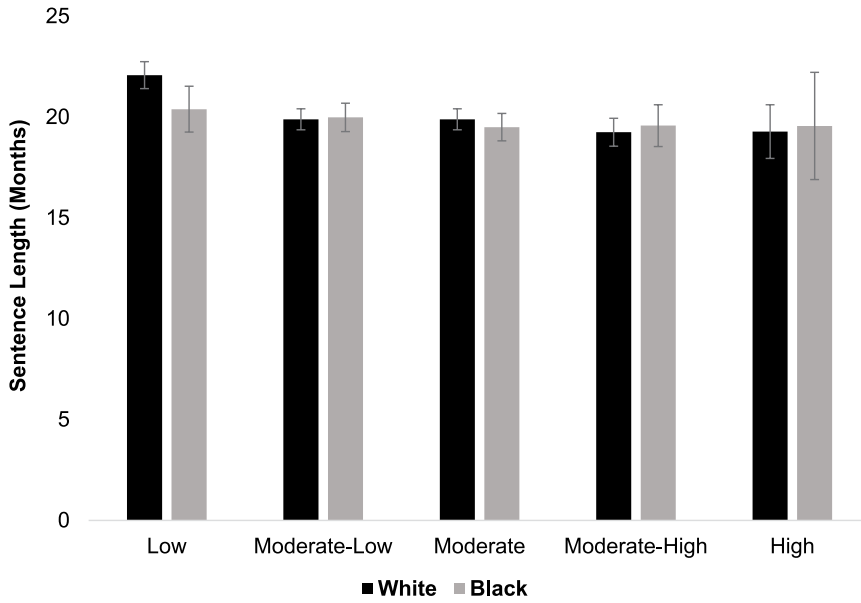

Figure 1: Sentence Length by LSI-R Risk Classification and Race
Note. Error bars reflect 99% confidence interval. LSI-R = Level of Service Inventory–Revised.

Sentence Length

As shown in Table 1, inclusion of predictors resulted in reduced between-probationer residual variability relative to the unconditional null model, suggesting that predictors accounted for 32.4% of Level 1 variability in sentence length. In Block 1, LSI-R risk bin classifications were negatively associated with sentence length, such that probationers with higher risk classifications had shorter sentences relative to probationers with Low risk classifications. However, sentence lengths did not differ significantly between White and Black probationers (p = .101). The addition of race by risk classification interaction terms in Block 2 resulted in a significant improvement in model fit relative to Block 1, p = .004. Specifically, we observed evidence of race by risk classification interactions for Low-Moderate (p < .001), Moderate (p = .009), and Moderate-High (p < .001) risk classifications relative to Low risk.

Decomposition of these interactions suggested these trends were largely driven by White probationers, who had significantly longer sentences when classified at Low risk, $F(1, 11{,}583) = 14.13$, $p < .001$, $d = 0.20$, relative to Black probationers and significant differences in sentence length across risk classifications, $F(4, 11{,}581) = 29.63$, $p < .001$, $d = 0.34$. In contrast, there were no between-group differences in sentence length across all other risk classifications, ps ⩾ .080 and no difference in sentence length across risk classifications for Black probationers, p = .196. These results are depicted in Figure 1, which shows the trend of shorter sentences associated with higher risk classifications was largely driven by White probationers, who received the longest sentence when classified at Low risk (M = 22.08, SE = 0.26), followed by Low-Moderate (M = 19.89, SE = 0.20) and Moderate (M = 19.89, SE = 0.20) classifications and shorter sentences at Moderate-High (M = 19.25, SE = 0.27) and

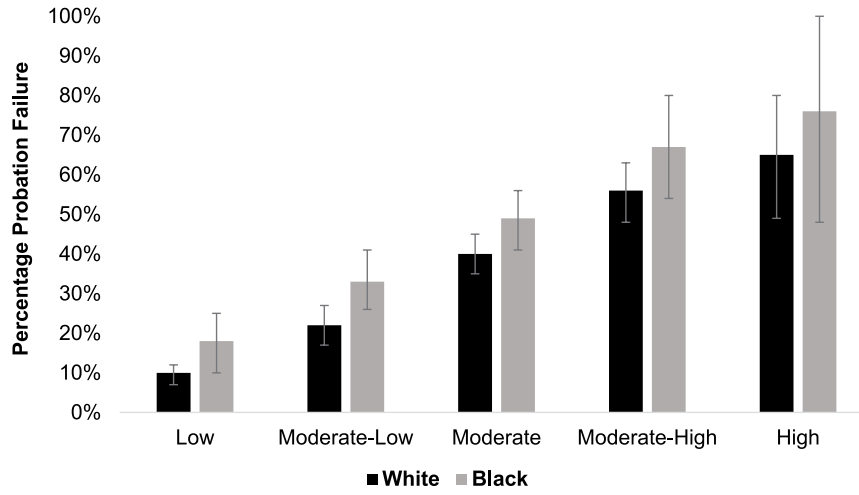

Figure 2: Percentage of Probationers With Probation Failure by LSI-R Risk Classification and Race
Note. Error bars reflect 99% confidence interval. LSI-R = Level of Service Inventory–Revised.

High (M = 19.28, SE = 0.52) risk classifications. In contrast, this trend was less pronounced among Black probationers, who received the shortest sentences when classified at Moderate risk (M = 19.50, SE = 0.27) and longer sentences at Moderate-High (M = 19.58, SE = 0.40) and High (M = 19.56, SE = 1.03) risk classifications. However, there was uncertainty around estimates at High risk classifications for both White and Black probationers, as illustrated by large confidence intervals in Figure 1. Furthermore, similar to White probationers, the average sentence length for Black probationers was longest at Low (M = 20.39, SE = 0.44) and Low-Moderate (M = 19.99, SE = 0.27) risk classifications.

As shown in Table 2, inclusion of predictors together with LSI-R total scores similarly explained a substantial portion of between-probationer variability in sentence length (32.2%). In Block 2, a significant LSI-R total score by race interaction effect contributed to a significant improvement in model fit over Block 1, p = .007. A decomposition of this interaction at +1 and −1 standard deviation from mean levels of LSI-R total scores showed White probationers had longer sentence lengths at Low risk levels (M = 21.88, SE = 0.19) relative to Black probationers (M = 21.24, SE = 0.26), p < .001, d = 0.05. Conversely, there was little difference in average sentence lengths at High risk levels between White (M = 20.64, SE = 0.20) and Black (M = 20.75, SE = 0.26) probationers, p = .298. In addition, although higher LSI-R total scores were associated with shorter sentence lengths for White probationers (p < .001, d = 0.10), a similar trend was not supported for Black probationers (p = .042).

Probation Failure

In Block 1, risk bin classifications incrementally predicted failure to complete probation, ps < .001, relative to reference levels (see Table 1). Furthermore, Black probationers were more likely to be terminated from probation without successful completion relative to White probationers, p < .001. In Block 2, however, we found limited support for race by risk

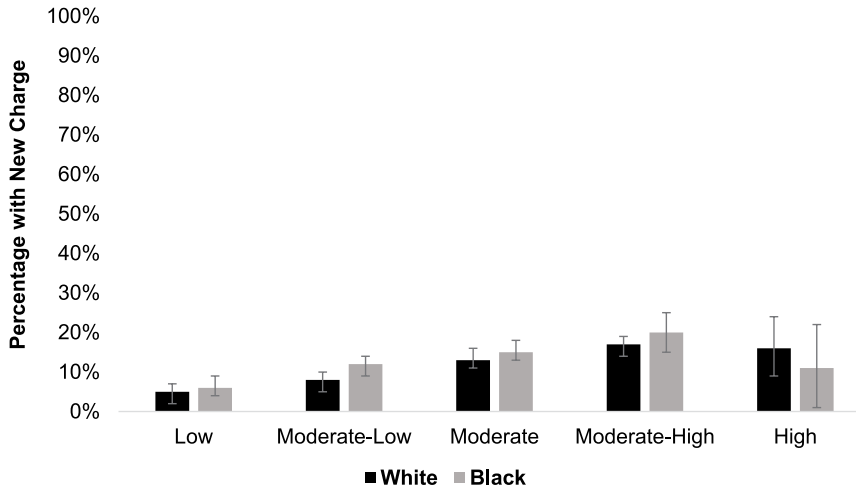

Figure 3: Percentage of Probationers With a New Charge by LSI-R Risk Classification and Race
Note. Error bars reflect 99% confidence interval. LSI-R = Level of Service Inventory–Revised.

classification interaction effects (*p*s ⩾ .263), as evidenced by a nonsignificant improvement in model fit over Block 1 ($p = .420$). This overall nonsignificant race by risk classification effect is depicted in Figure 2. As shown, although the rate of offending associated with each risk classification was higher for Black probationers relative to White probationers, the increase in offending risk from lower to higher risk classifications was proportional for both groups. This suggests that assessments produced similar predictive accuracy for both White and Black probationers across risk classifications. We observed similar results in Table 2, including a nonsignificant LSI-R total score by race interaction in Block 2 ($p = .222$) and corresponding negligible improvement in model fit, $p = .171$.

New Charge

In Block 1, LSI-R risk classifications again showed incremental prediction of any new charge relative to reference levels, *p*s < .001 (see Table 1). Black probationers had a slightly elevated odds of acquiring a new charge relative to White probationers, $p < .001$. In Block 2, we found limited evidence of race by risk classification interactions, *p*s ⩾ .119 and a nonsignificant improvement in model fit, $p = .096$. As shown in Figure 3, increases in offending rates were relatively proportional between Black and White probationers from Low risk classification to Moderate-High risk. Although average rates of offending were lower for Black probationers classified at High risk relative to White probationers, variability around these estimates limited any conclusive inferences. Results for models involving LSI-R total scores were similar (see Table 2). In Block 2, there was no evidence of a race by LSI-R total score interaction, $p = .121$, and improvement in model fit over Block 1 was nonsignificant, $p = .142$.

FIVE LEVEL POST HOC COMPARISONS

To further our examination of disparate impact, we examined whether application of a five-level approach based on predicted probabilities of offending would alter the

distribution of risk classifications for Black and White probationers. Results showed most probationers maintained the same risk classification (n = 5,928, 50.3%) or increased one level in risk classification (n = 5,350, 45.4%). Fewer participants decreased by one level in risk classification (n = 223, 1.9%) or increased in risk by two levels (n = 273, 2.3%). The vast majority of participants were classified in Level III (n = 9,106, 77.3%), with fewer participants classified in Level I (n = 41, 0.3%), Level II (n = 1,087, 9.2%), or Level IV (n = 1,540, 13.1%). No participants were classified in Level V. Results of a cross-tabulation suggested a small but significant difference in change in risk levels as a function of race, $\chi^2(3) = 14.96$, $p = .002$, $\Phi = .04$. Specifically, Black probationers were less likely to have a one-level decrease in risk classification relative to White probationers (1.4% and 2.1%, respectively). Black probationers were also less likely to have a two-level increase in risk classification relative to White probationers (1.7% and 2.5%, respectively). However, Black probationers were more likely to have no change in risk classification relative to White probationers (52.3% and 49.7%, respectively). There were no differences between Black and White probationers in having a one-level change in risk classification (44.6% and 45.7%, respectively).

DISCUSSION

Risk assessment instruments are now implemented in correctional settings across the United States as an evidence-based strategy to inform sentencing and other correctional decision-making. Despite the widespread use of such instruments, critics—including former U.S. Attorney General Eric Holder—have cautioned that their reliance on static risk factors may exacerbate racial inequality by biasing racial minorities toward higher risk classifications. A growing number of studies have investigated the potential for racial disparities in the predictive validity of risk assessments conducted in correctional contexts; however, few studies have examined the potential for racial bias in the application of risk assessments in sentencing decisions, especially in the United States. We addressed this gap by investigating potential for racial bias in the use of the LSI-R, a commonly employed risk assessment instrument in correctional settings, to inform sentence length as well as predict key probation outcomes in a state-wide sample of Black and White probationers. Our findings suggested potential for racial bias in the use of LSI-R assessments to inform sentencing decisions, though effects were small. Furthermore, we found little evidence of bias in the ability of LSI-R assessments to predict community outcomes similarly for White and Black probationers. Below, we outline and discuss findings in greater detail.

INTEGRATION OF STUDY FINDINGS

Our findings showed that LSI-R risk assessment scores and risk classifications negatively predicted sentence length. Although counterintuitive, this finding has been reported in similar investigations of risk assessments used in sentencing decisions as a byproduct of high-risk defendants being more likely to commit minor crimes and have shorter sentences relative to low-risk defendants (van Wingerden, van Wilsem, & Moerings, 2014). Our findings suggested this trend was driven primarily by White probationers at low-risk levels, who received sentences that were on average 2 months longer relative to those received by Black probationers classified at Low risk levels. Consistent with our investigation of disparate impact, our findings raise the possibility of dissimilar application of risk assessment findings by race, particularly when defendants are classified at Low risk. However, this

disparate application appears to be both in the favor of Black probationers, who received shorter sentences at Low risk levels relative to White probationers and comparable sentences at High risk levels, and fairly small in effect size. Although sentences for Black probationers were slightly longer at Moderate-High and High risk classifications, they were not significantly so, likely attributable to greater variability in sentence lengths for Black probationers classified at these levels.

Because our investigation examined predictive associations, we are limited in our ability to fully explain these trends or infer risk assessment information as the causal factor in these differences. However, because we controlled for offense severity per Kansas sentencing guidelines, our findings could reflect the discretion of judges to work within a specific range of sentencing possibilities. Accordingly, judges may have been more likely to assign blame to White defendants classified at Low risk levels relative to Black defendants and thus less likely to rely on risk assessment results to adjudicate cases. Alternatively, there could have been other characteristics (e.g., legal, extralegal) of White defendants unmeasured in this study that may have contributed to judicial perceptions of heightened risk and resulted in harsher sentences. Accordingly, lower risk ratings may have reflected the relative social advantage of White defendants based on risk factors measured by the LSI-R. However, these explanations are purely speculative and warrant investigation in future research.

Still unclear in the broader literature is how information generated by risk assessments is used in correctional practice to inform supervision and sentencing decisions. For example, in the context of sentencing, state laws regarding the use of risk assessments often mandate only their use in specific contexts or with specific groups of offenders and are less focused on establishing clear decision-making thresholds (Monahan & Skeem, 2016; Widgery, 2015). Compounding this issue are differences in the interpretation of risk categories and probabilities of offending across risk assessment instruments, which have the potential to obscure the meaning of risk labels such as “low,’ “moderate,” or “high” (Hanson et al., 2017). In the absence of clear decision-making guidelines for how information generated from risk assessments should either mitigate or aggravate sentencing and supervision decisions, there remains potential for disproportionate application of risk assessment results.

In contrast to sentencing decisions, we found little evidence of racial bias in the ability of LSI-R assessments to predict probation failure or a new charge. Although the LSI-R was originally developed for probation and parole populations (Andrews et al., 2010), few studies have examined the predictive validity of LSI-R assessments with respect to probation outcomes, let alone between racial groups. Consistent with our definition of predictive bias, risk classifications and total scores produced similar levels of predictive accuracy between White and Black probationers. We observed these results even with generally higher base rates of probation failure or any new charge for Black probationers, which both supports the need for distinct definitions of fairness and illustrates the impossibility of satisfying all definitions of fairness simultaneously (Berk et al., 2017; Kleinberg et al., 2016). Our findings run contrary to previously reported findings on racial differences in the predictive validity of LSI-R assessments (e.g., Chenane et al., 2015; Ostermann & Salerno, 2016) but are similar to investigations conducted on risk assessments from other instruments (Skeem & Lowenkamp, 2016). Of note, few studies have explicitly modeled a risk classification by race interaction (Lowder et al., 2017; Skeem & Lowenkamp, 2016), instead relying on descriptive comparisons of predictive validity estimates (such as Area Under the Curve

[AUC] values) by group (e.g., Ostermann & Salerno, 2016). However, this approach relies on total scores instead of risk estimates, which are more likely to be used in practice, and often fails to provide a statistical significance indicator for differences in predictive validity estimates between groups. Thus, varying analytical approaches, often compounded by imprecise definitions of fairness, limit the ability to contextualize findings in the existing literature.

Finally, although the justification for a five-level approach to the operationalization of offending risk across instruments is consistent with the need to clarify and standardize how risk assessments are used to inform correctional decisions, we found little evidence that such an approach would alter the distribution of risk classifications for minority offenders. In contrast, White probationers were more like to experience changes in risk classification, including a one-level reduction and a two-level increase, relative to Black probationers. To be sure, our findings do not inform whether a five-level approach would reduce racial bias in decision-making, or whether five-level categories would produce different predictive validity estimates across racial groups relative to original risk categories. We relied on a base rate of offending computed from a population of Kansas probationers; however, the recommended offending probabilities for the five-level approach are still in development (Hanson et al., 2017), and thus, there will be need for further research on this approach and its potential for standardizing the interpretation of risk categories across instruments once these probabilities have been established.

At this point, a small body of evidence indicates potential for racial bias in the predictive accuracy and application of risk assessments. However, whether risk assessments are less biased relative to unstructured approaches alone is a separate and arguably more significant question. In clinical contexts, structured assessments have been shown to produce less biased and more accurate assessments relative to unstructured approaches (Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Current sentencing guidelines are largely based on crime type and crime severity, and many sentencing grids additionally incorporate criminal history into sentencing guidelines (Frase, 2005). Thus, risk assessments have the potential to serve a mitigating role in standard sentencing guidelines, which are overwhelmingly based on indicators of crime and criminal involvement for which we know minorities are at a relative disadvantage. In particular, there is evidence to suggest that instruments measuring protective factors may be less biased toward racial minorities because they focus on measuring positive aspects of offenders' lives that may buffer against risk factors and decrease the likelihood of experiencing adverse outcomes (Lowder et al., 2017; Perrault, Vincent, & Guy, 2017). Particularly, in high-risk offenders, protective factors may more clearly delineate recidivism risk among offenders presenting with similar risk factors (de Ruiter & Nicholls, 2011). Indeed, by focusing less on criminal history and other static factors, risk assessments have the potential to be less racially biased (Skeem & Lowenkamp, 2016).

LIMITATIONS AND FUTURE DIRECTIONS

Our findings should be considered in light of several limitations, which may inform future research at the nexus of race, risk assessment, and criminal sentencing. First, we did not have access to item-level data to calculate internal consistency nor did we have information on inter-rater reliability, beyond staff training procedures. However, the omission of

reliability estimates, particularly inter-rater reliability, is not altogether uncommon in risk assessment research in criminal justice settings (Desmarais et al., 2016). Second, our sample—although large and state-wide—was restricted to defendants who were sentenced to probation only, excluding probationers who may have additionally completed a prison sentence. This decision was made to increase the internal validity of findings by removing prison sentencing as a potential confound, particularly given that Kansas state policy requires presentencing LSI-R assessments to inform probation placement and sentencing, not prison sentencing. We acknowledge that this decision came at the cost of reduced generalizability of our sample.

Third, beyond statutory requirements to include the LSI-R as part of the presentence investigation to inform sentencing and supervision placement, we had limited information on how presentencing assessments were utilized in sentencing procedures. Utilization of risk assessments likely differed by jurisdiction and even by judge. However, this limitation raises a broader question regarding how risk assessment results should be incorporated into sentencing guidelines to promote the best outcomes. The current use of risk assessment to inform sentencing decisions most often follows a hybrid theory of limiting retributivism, whereby risk assessments may serve to alter sentence length within specific bounds set by the severity of the crime (Monahan & Skeem, 2016). Much of the application appears to be at the discretion of the judge, which introduces greater variability in sentencing decisions and the possibility of racially disparate decisions. To date, there have been few investigations on judicial attitudes toward risk assessment or the use of risk assessment information in judicial decision-making (Cole, 2007; Hyatt & Chanenson, 2016). Additional research into how risk assessments are incorporated into correctional decision-making may help inform opportunities for further training or introduction of structured guidelines for use of risk assessment results.

Third, and relatedly, our findings were observational. We did not experimentally manipulate key variables such as risk level and race to examine effects on sentencing decisions. We attempted to control for potential third variables influencing sentencing decisions and probation outcomes such as level of charge, crime type, and personal characteristics. We were unable to control for other legal or extralegal factors that have been found to predict sentencing outcomes more broadly (Hart, Miethe, & Regoeczi, 2014; Johnson, 2006; Ulmer & Johnson, 2006). Furthermore, we were unable to speak to sentencing decisions made in the absence of risk assessments. Like the present investigation, most studies of race and risk assessment have been retrospective in nature, reflecting the challenge of implementing new risk assessment protocols when many states are already required by state law to use a risk assessment to inform correctional decision-making. In fact, we are only aware of one recent investigation that employed a quasi-experimental design with a propensity score-matched comparison group to examine the effect of presentencing risk assessments on sentencing outcomes (van Wingerden et al., 2014). In this study, defendants who received a presentencing assessment were matched to defendants whose cases were adjudicated during the same period without a presentencing assessment. However, this study was conducted outside of the United States and did not examine the disparate impact of risk assessments with respect to race or any other protected class.

Moving beyond the question of whether instruments produce comparable predictive validity estimates and toward addressing whether correctional decisions are made more or less racially disparate by the use of risk assessments will require more sophisticated research

designs. This is an essential question and one that the current scholarship on fairness in risk assessment often neglects. Risk assessments implemented in correctional contexts should be contextualized as an alternative to practice as usual. Decades of research have documented the realities of racial bias and disparities in sentencing and other correctional decisions (e.g., Abrams et al., 2012; Bales & Piquero, 2012; Kutateladze et al., 2014; Sweeney & Haney, 1992; Wooldredge, 2012; Wu, 2016). Risk assessments have the potential to reduce these disparities by providing a structure for the evaluation of future offending risk within the scope of law and framework of limiting retributivism (Monahan & Skeem, 2016). However, there is need for more prospective and carefully controlled investigations on risk assessments as an alternative to practice as usual to cement risk assessments as a truly evidence-based practice capable of improving the equitable and effective administration of justice.

CONCLUSION

Despite the proliferation of risk assessments in correctional contexts, there have been few comparative studies examining whether risk assessments contribute to improved risk management strategies relative to decisions in the absence of risk assessments. These investigations are key not only to establish risk assessments as an evidence-based practice with the potential to reduce the number of adults under correctional supervision but also to address directly the potential for racially biased application of risk assessment findings. Our findings suggest the potential for disparate application of risk assessment results to probation sentencing decisions. However, it remains to be seen whether these findings generalize to other jurisdictions, assessments completed with other instruments, and other forms of correctional decision-making. Given the widespread use of risk assessments, such research is imperative to address growing and legitimate concerns about the potential for racially biased decision-making in correctional practice.

ORCID ID

Evan M. Lowder https://orcid.org/0000-0002-5855-2479

REFERENCES

Abrams, D. S., Bertrand, M., & Mullainathan, S. (2012). Do judges vary in their treatment of race? *The Journal of Legal Studies*, *41*, 347-383. doi:10.1086/666006

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341-382. doi:10.1177/0011000005285875

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory–Revised: User's manual*. Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2001). *Level of Service Inventory–Revised (LSI-R): User's manual*. Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law*, *16*, 39-55. doi:10.1037/a0018362

Andrews, D. A., Bonta, J. L., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, *17*, 19-52. doi:10.1177/0093854890017001004

Andrews, D. A., Bonta, J. L., & Wormith, S. (2010). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto & K. S. Douglas (Eds.), *Handbook of violence risk assessment* (1st ed., pp. 199-225). New York, NY: Routledge.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, *28*, 369-404. doi:10.1111/j.1745-9125.1990.tb01330.x

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Babschishin, K. M., Kroner, D. G., & Hanson, R. K. (2017). Standardized risk/need levels for corrections. *Crime Scene*, *24*(1), 9-13.

Bales, W. D., & Piquero, A. R. (2012). Racial/ethnic differentials in sentencing to incarceration. *Justice Quarterly*, *29*, 742-773. doi:10.1080/07418825.2012.659674

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. *Arxiv*. Retrieved from https://arxiv.org/abs/1703.09207

Callis, R. R., & Kresin, M. (2016). *Residential vacancies and homeownership in the second quarter*. Washington, DC: U.S. Census Bureau, U.S. Department of Commerce.

Casey, P. M., Elek, J. K., Holt, K. A., Johnson, T. D., Miller, S. S., & Warren, R. K. (2013). *Use of risk and needs assessment information at sentencing: 7th Judicial District, Idaho*. Williamsburg, VA: Center for Sentencing Initiatives, Research Division, National Center for State Courts.

Chenane, J. L., Brennan, P. K., Steiner, B., & Ellison, J. M. (2015). Racial and ethnic differences in the predictive validity of the Level of Service Inventory–Revised among prison inmates. *Criminal Justice and Behavior*, *42*, 286-303. doi:10.1177/0093854814548195

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Routledge.

Cole, J. D. P. (2007). The umpires strike back: Canadian judicial experience with risk-assessment instruments. *Canadian Journal of Criminology & Criminal Justice*, *49*, 493-517. doi:10.3138/cjccj.49.4.493

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 797-806). New York, NY: Association for Computing Machinery. doi:10.1145/3097983.309809

de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *The International Journal of Forensic Mental Health*, *10*, 160-170. doi:10.1080/14999013.2011.600602

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services*, *13*, 206-222. doi:10.1037/ser0000075

Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice and Behavior*, *35*, 1095-1108. doi:10.1177/0093854808320497

Frase, R. S. (2005). State sentencing guidelines: Diversity, consensus, and unresolved policy issues. *Columbia Law Review*, *105*, 1190-1232.

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, *102*, 813-823. doi:10.1198/016214506000001040

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19-30.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, *26*, 441-462. doi:10.1146/annurev.soc.26.1.441

Hanson, K. R., Bourgon, G., McGrath, R. J., Kroner, D., D'Amora, D. A., Thomas, S. S., & Tavarez, L. P. (2017). *A five-level risk and needs system: Maximizing assessment results in corrections through the development of a common language*. New York, NY: Council of State Governments Justice Center and National Reentry Resource Center.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, *27*, 237-243. doi:10.1525/fsr.2015.27.4.237

Hart, T. C., Miethe, T. D., & Regoeczi, W. C. (2014). Contextualizing sentencing disparities: Using conjunctive analysis of case configurations to identify patterns of variability. *Criminal Justice Studies*, *27*, 344-361. doi:10.1080/1478601X.2014.947031

Heck, R. H., & Thomas, S. L. (2008). *An introduction to multilevel modeling techniques* (2nd ed.). New York, NY: Routledge.

Helmus, L., & Hanson, K. R. (2011). More fun with statistics! How to use logistic regression to predict criminal recidivism risk. *Crime Scene*, *18*(2), 8-12.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70.

Hox, J. J. (1994). *Applied multilevel analysis*. Amsterdam, The Netherlands: TT-Publikaties.

Hox, J. J., Moerbeek, M., & van der Schoot, R. (2010). *Multilevel analysis: Techniques and applications, second edition* (2nd ed.). New York, NY: Routledge.

Hyatt, J., & Chanenson, S. L. (2016). *The use of risk assessment at sentencing: Implications for research and policy* (SSRN Scholarly Paper No. ID 2961288). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2961288

Johnson, B. D. (2006). The multilevel context of criminal sentencing: Integrating judge-and county-level influences. *Criminology*, *44*, 259-298. doi:10.1111/j.1745-9125.2006.00049.x

Kaeble, D., & Bonczar, T. P. (2016). *Probation and parole in the United States, 2015*. Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv*. Retrieved from https://arxiv.org/abs/1609.05807

Kutateladze, B. L., Andiloro, N. R., Johnson, B. D., & Spohn, C. C. (2014). Cumulative disadvantage: Examining racial and ethnic disparity in prosecution and sentencing. *Criminology*, *52*, 514-551. doi:10.1111/1745-9125.12047

Lawrence, A. (2013). *Trends in sentencing and corrections: State legislation*. Denver, CO: National Conference of State Legislatures.

Lowder, E. M., Desmarais, S. L., Rade, C. B., Johnson, K. L., & Van Dorn, R. A. (2017). Reliability and validity of START and LSI-R assessments in mental health jail diversion clients. *Assessment*. Advance online publication. doi:10.1177/1073191117704505

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E. J. (2004). Assessing the inter-rater agreement of the Level of Service Inventory–Revised. *Federal Probation*, *68*(3), 34-46.

Lowenkamp, C. T., Lovins, B., & Latessa, E. J. (2009). Validating the Level of Service Inventory–Revised and the Level of Service Inventory: Screening Version with a sample of probationers. *The Prison Journal*, *89*, 192-204. doi:10.1177/0032885509334755

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology*, *12*, 489-513. doi:10.1146/annurev-clinpsy-021815-092945

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*, 771-785. doi:10.1177/0146167201277001

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2014). Thirty years of research on the Level of Service scales: A meta-analytic examination of predictive accuracy and sources of variability. *Psychological Assessment*, *26*, 156-176. doi:10.1037/a0035080

Ostermann, M., & Salerno, L. M. (2016). The validity of the Level of Service Inventory–Revised at the intersection of race and gender. *The Prison Journal*, *96*, 554-575. doi:10.1177/0032885516650878

Perrault, R. T., Vincent, G. M., & Guy, L. S. (2017). Are risk assessments racially biased? Field study of the SAVRY and YLS/CMI in probation. *Psychological Assessment*, *29*, 664-678. doi:10.1037/pas0000445

Proctor, B. D., Semega, J. L., & Kollar, M. A. (2016). *Income and poverty in the United States: 2015*. Washington, DC: U.S. Census Bureau, U.S. Department of Commerce. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2016/demo/p60-256.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: SAGE.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3-15. doi:10.1177/096228029900800102

Schlager, M. D., & Simourd, D. J. (2007). Validity of the Level of Service Inventory–Revised (LSI-R) among African American and Hispanic male offenders. *Criminal Justice and Behavior*, *34*, 545-554. doi:10.1177/0093854806296039

Simourd, D. (2006). *Validation of risk/needs assessments in the Pennsylvania Department of Corrections*. Hampden Township, PA: Department of Corrections.

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences & the Law*, *31*, 55-73. doi:10.1002/bsl.2053

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*, 499-513. doi:10.1016/j.cpr.2010.11.009

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*, 680-712. doi:10.1111/1745-9125.12123

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, Article 803.

Sweeney, L. T., & Haney, C. (1992). The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, *10*, 179-195. doi:10.1002/bsl.2370100204

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston, MA: Pearson Education.

Ulmer, J. T., & Johnson, B. (2006). Sentencing in context: A multilevel analysis. *Criminology*, *42*, 137-178. doi:10.1111/j.1745-9125.2004.tb00516.x

U.S. Census Bureau. (2017a). *Table 1. Educational attainment of the population 18 years and over, by age, sex, race, and Hispanic origin: 2016*. Retrieved from https://www.census.gov/data/tables/2016/demo/education-attainment/cps-detailed-tables.html

U.S. Census Bureau. (2017b). *Table 1. Median value of assets for households, by type of asset owned and selected characteristics: 2013*. Retrieved from https://www.census.gov/data/tables/2013/demo/wealth/wealth-asset-ownership.html

U.S. Department of Justice. (2014, August 1). *Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th annual meeting*. Retrieved from http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

van Eijk, G. (2016). Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, *19*, 463-481. doi:10.1177/1462474516666282

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.

Vornovitsky, M., Gottschalck, A., & Smith, A. (2011). *Distribution of household wealth in the U.S.: 2000 to 2011*. Washington, DC: U.S. Census Bureau. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2011/demo/wealth-distribution-2000-to-2011.pdf

Vose, B., Cullen, F. T., & Smith, P. (2008). The empirical status of the Level of Service Inventory. *Federal Probation*, *72*(3), 22-29.

Whiteacre, K. W. (2006). Testing the Level of Service Inventory–Revised (LSI-R) for racial/ethnic bias. *Criminal Justice Policy Review*, *17*, 330-342. doi:10.1177/0887403405284766

Widgery, A. (2015). *Trends in pretrial release: State legislation*. Denver, CO: National Conference on State Legislatures.

Wooldredge, J. (2012). Distinguishing race effects on pre-trial release and sentencing decisions. *Justice Quarterly*, *29*, 41-75. doi:10.1080/07418825.2011.559480

Wu, J. (2016). Racial/ethnic discrimination and prosecution: A meta-analysis. *Criminal Justice and Behavior*, *43*, 437-458. doi:10.1177/0093854815628026

Evan M. Lowder, PhD, is a postdoctoral research associate in the School of Public and Environmental Affairs at Indiana University Purdue University Indianapolis. Her research focuses on strategies to reduce offending and improve behavioral health outcomes among justice-involved adults, with specific focus on risk and needs assessment.

Megan M. Morrison, PhD, is an assistant professor of psychology at Tennessee State University. Her research focuses on racial stereotyping, prejudice, and discrimination with emphasis on intercultural relationships and multicultural individuals.

Daryl G. Kroner, PhD, is a professor of criminology and criminal justice at Southern Illinois University at Carbondale. His research focuses on mental health assessment of female offenders, dynamic risk assessment during community supervision, evaluation of community interventions, and examination of treatment attrition predictors.

Sarah L. Desmarais, PhD, is an associate professor of psychology and coordinator of the Applied Social and Community Psychology Graduate Program at North Carolina State University. Her research focuses on the assessment and treatment of risks and needs associated with criminal behavior, interpersonal violence, and terrorism.