False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."

Anthony W. Flores
California State University, Bakersfield
Kristin Bechtel
Crime and Justice Institute at CRJ
Christopher T. Lowenkamp
Administrative Office of the United States Courts
Probation and Pretrial Services Office

The validity and intellectual honesty of conducting and reporting analysis are critical, since the ramifications of published data, accurate or misleading, may have consequences for years to come.
—Marco and Larkin, 2000, p. 692

PROPUBLICA RECENTLY RELEASED a much-heralded investigative report claiming that a risk assessment tool (known as the COMPAS) used in criminal justice is biased against black defendants.[1][2] The report heavily implied that such bias is inherent in all actuarial risk assessment instruments (ARAIs).

We think ProPublica's report was based on faulty statistics and data analysis, and that the report failed to show that the COMPAS itself is racially biased, let alone that other risk instruments are biased. Not only do ProPublica's results contradict several comprehensive existing studies concluding that actuarial risk can be predicted free of racial and/or gender bias, a correct analysis of the underlying data (which we provide below) sharply undermines ProPublica's approach.

Our reasons for writing are simple. It might be that the existing justice system is biased against poor minorities due to a wide variety of reasons (including economic factors, policing patterns, prosecutorial behavior, and judicial biases), and therefore, regardless of the degree of bias, risk assessment tools informed by objective data can help *reduce* racial bias from its current level. It would be a shame if policymakers mistakenly thought that risk assessment tools were somehow worse than the status quo. Because we are at a time in history when there appears to be bipartisan political support for criminal justice reform, one poorly executed study that makes such absolute claims of bias should not go unchallenged. The gravity of this study's erroneous conclusions is exacerbated by the large-market outlet in which it was published (ProPublica).

Before we expand further into our criticisms of the ProPublica piece, we describe some context and characteristics of the American criminal justice system and risk assessments.

Mass Incarceration and ARAIs

The United States is clearly the worldwide leader in imprisonment. The prison population in the United States has declined by small percentages in recent years and at year-end 2014 the prison population was the smallest it had been since 2004. Yet, we still incarcerated 1,561,500 individuals in federal and state correctional facilities (Carson, 2015). By sheer numbers, or rates per 100,000 inhabitants, the United States incarcerates more people than just about any country in the world that reports reliable incarceration statistics (Wagner & Walsh, 2016).

Further, it appears that there is a fair amount of racial disproportion when comparing the composition of the general population with the composition of the prison population. The 2014 United States Census population projection estimates that, across the U.S., the racial breakdown of the 318 million residents comprised 62.1 percent white, 13.2 percent black or African American, and 17.4 percent Hispanic. In comparison, 37 percent of the prison population was categorized as black, 32 percent was categorized as white, and 22 percent as Hispanic (Carson, 2015). Carson (2015:15) states that, "As a percentage of residents of all ages at yearend 2014, 2.7 percent of black males (or 2,724 per 100,000 black male residents) and 1.1 percent of Hispanic males (1,090 per 100,000 Hispanic males) were serving sentences of at least 1 year in prison, compared to less than 0.5 percent of white males (465 per 100,000 white male residents)."

Aside from the negative effects caused by

[1] The authors wish to thank James Bonta, Francis Cullen, Edward Latessa, John Monahan, Ralph Serin, Jennifer Skeem, and Stuart Buck for their thoughtful comments and suggestions.

[2] The main article and an accompanying analysis report were authored by the same individuals, albeit with a different order of authorship. The main ProPublica article is cited as Angwin, Larson, Mattu, and Kirchner (2016) or Angwin et al. (2016). The analysis report is cited as Larson et al. (2016).

imprisonment, there is a massive financial cost that extends beyond official correctional budgets. A recent report by The Vera Institute of Justice (Henrichson & Delaney, 2012) indicated that the cost of prison operations (including such things as pension and insurance contributions, capital costs, legal fees, and administrative fees) in 40 states participating in their study was 39.5 billion (with a b) dollars per year. The financial and human costs, and perhaps absurdity, of these practices have become so obvious that there has been bipartisan support for efforts to develop solutions to reduce the amount of money spent on incarceration and the number of lives negatively impacted by incarceration (Skeem & Lowenkamp, 2016b).

An example of one such effort has been the investigation of the use of ARAIs to partially inform decisions related to sentencing and other correctional decisions. Whether it is appropriate to use ARAIs in criminal justice settings is a popular debate. However, as Imrey and Dawid (2015:18)[3] note, the debates and "… considerations [of using ARAIs in such settings] are properly functions of social policy, not statistical inference." That is, there might be much to debate about how and why we would *use* valid ARAIs. The issue that is no longer up for debate is that ARAIs predict outcomes more strongly and accurately than professional judgment alone. Several studies and meta-analyses have reached similar conclusions indicating that actuarial risk assessments are superior to unstructured professional judgment in terms of predicting the likelihood of both general recidivism and even specific recidivism outcomes (Grove, Zald, Lebow, Snitz, & Nelson, 2000), including future sex offending (Hanson & Morton-Bourgon, 2009). Noteworthy research on the predictive accuracy of risk assessments can be attributed to Meehl (1954) and Grove et al. (2000), including the oft-cited and comprehensive review of risk assessments from Andrews, Bonta, and Wormith (2006).

Given that this research often goes unrecognized by those concluding that ARAIs cannot be relied upon to predict outcomes, it is relevant to clarify what the potential consequences are for ignoring (presumably unintentionally) a vast body of research on the performance of ARAIs. Specifically, the implications could be as serious as dismissing the use of risk assessments outright. This type of abrupt response and return to subjective judgment would be unethical, and one poorly informed statement should not replace over 60 years of research in which consistent findings are produced in support of ARAIs.

ARAIs are intended to inform objective decision-making, so proper administration of the instrument and clear guidance on what information risk assessments are capable of reliably providing for a target population are relevant points of discussion. What is equally important is that the development of these tools be rigorous and that subsequent tests of their performance in predicting recidivism include independent evaluations. Finally, critiques of risk assessments, including questions about racial bias, should be properly conducted and described. Thankfully, there are empirical standards for testing whether assessments are biased—standards that were not discussed or applied in the ProPublica pieces.

One of the more common concerns that arise in the discourse on the use of risk assessment in correctional and sentencing contexts is racial bias. Given the racial disproportionality already seen in prison populations (and at other points in the criminal justice process), racial bias is a salient issue for the use of ARAIs or any other method to structure decision-making. But concerns that the use of ARAIs would increase racial disproportionality were drawn from *hypothetical* or *theoretical* linkages and limited empirical evidence between certain risk factors and race (see Skeem & Lowenkamp, 2016a). It is unfortunate that this concern—race-based bias in risk assessment—is threatening to stall sentencing and correctional reform, especially when it is likely, given the racial disproportionality in the correctional system, that minorities could benefit most from unwinding mass incarceration. Still, these concerns over bias *are legitimate*. At the same time, these concerns can and should be *properly* investigated.

In their attempt to investigate test bias of the Northpointe COMPAS across different categories of race, the ProPublica authors constructed four multivariate models. Two models predicted the likelihood that the defendant was classified as high-risk and two estimated the effect of race on the relationship between the COMPAS score and recidivism (any arrest and arrest for a violent offense). The authors conclude that

> The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.
>
> When a full range of crimes were taken into account—including misdemeanors such as driving with an expired license—the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.
>
> We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.
>
> The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants.

We appreciate that Angwin et al. (2016) made their data available for subsequent analyses by other researchers but take issue with how they analyzed the data and, consequently, their conclusions. Before we proceed further, we want to make it clear that we are not supporting or endorsing the idea of using risk assessment at sentencing (although we do support its use at certain decision points in the correctional system) nor are we advocating for the Northpointe COMPAS. We also are not making any blanket statements about race, test bias, and all ARAIs. With the previous qualifications, we present five concerns that we have with the analyses (Larson, Mattu, Kirchner, & Angwin, 2016) and the accompanying article by Angwin et al. (2016).

Criticisms of Angwin et al. (2016)

First, Angwin et al. (2016) conducted a study on a sample of pretrial defendants to determine if an instrument (the COMPAS) was biased when that instrument was not designed for use on pretrial defendants. Specifically, the COMPAS scales were developed upon and for individuals on post-disposition supervision. Further, the original sample for the ProPublica study also comprised probation and parole clients; however, Larson et al. (2016) excluded these relevant subjects from the study but failed to provide a detailed and acceptable reason for doing so. The sample they used (and shared for subsequent analysis) included only pretrial defendants, i.e., offenders who have not been convicted of the offenses for which they are being detained. This is a relevant distinction, as ARAIs that are intended to predict

[3] Also see Dawid, 2014, and Harris, Lowenkamp, & Hilton, 2015.

general and violent recidivism are typically developed and administered to probationers and parolees. Pretrial ARAIs are intended to predict different outcomes, such as failure to appear, for defendants. However, Larson et al. (2016) removed failure to appear arrests as an outcome measure for their analysis of pretrial defendants.

Additional clarification should be offered related to the COMPAS scales and their use in Broward County, Florida. The COMPAS does have a scale to examine pretrial failure outcomes, and Broward County does administer the pretrial, general recidivism, and violent recidivism scales to pretrial defendants; however, the general and violent recidivism scales are only appropriate for those on post-disposition supervision, when recidivism data would be collected within a specified time frame. The COMPAS validation study that the ProPublica authors cite to justify their definition and interpretation of their measures of recidivism (i.e., Brennan, Dieterich, & Ehret, 2009, p. 25) actually indicates that the COMPAS recidivism scales are intended to predict new offenses with probationer samples. There is no mention that the COMPAS recidivism scales are intended to predict recidivism for pretrial defendants (See page 25 from Brennan, Dieterich, & Ehret, 2009). Note, the purpose of the current study is not to address Broward County's use of the COMPAS scales with pretrial defendants, but we would strongly urge that examinations into the performance of an ARAI begin with a solid understanding of the tool's purpose, target population, and intended outcome(s).

Second, the authors force a dichotomy on the COMPAS. The COMPAS was not made to make absolute predictions about success or failure. Instead, it was designed to inform probabilities of reoffending across three categories of risk (low, medium, and high). Further, in their false positive/false negative analysis the authors collapsed all the moderate and high-risk defendants in the "high" category. The standard for this is to put all the moderate and high-risk defendants in a category and then reverse that and put low and moderate into a collapsed "low" category to observe if there are statistical changes as a result. See Singh (2013) for a methodological primer regarding performance indicators for ARAIs.

Third, the authors equate racial differences in mean scores on a risk assessment instrument (which would be highlighted by their model referenced in number 2 above) with test bias. This is not true—not true at all. See the Standards for Educational and Psychological Testing (discussed in the following point).

Fourth, well-established and accepted standards exist to test for bias in risk assessment. Larson et al. (2016) and Angwin et al. (2016) do not mention—or appear to be aware—that such standards exist. The analysis conducted in the ProPublica article fails to actually test for bias within these standards, which is critical given that this is the main focus of the report. Skeem and Lowenkamp (2016a) cover this issue extensively in their evaluation of the federal Post Conviction Risk Assessment (PCRA) and properly test for predictive bias within the guidelines from Standards for Educational and Psychological Testing. (For more information, see American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Fifth, Larson et al. (2016) overstate the effect of their results and fail to cite the limitations of their study. It is well known—and commonly taught in introductory statistics courses—that even trivial differences can attain statistical significance in large sample sizes. To address this, researchers have several options to select from, including pulling a random but smaller sample of cases from the original sample to conduct the analysis or setting the values to test for significance higher (e.g., $p. \leq 05$, $p \leq .01$, $p \leq .001$) as sample size increases. Larson et al. (2016) take the opposite approach: Even though the interaction terms in their two Cox Regression models do not reach statistical significance *by conventional standards* applied with relatively small samples, the authors interpret a significant difference when $p = 0.0578$. Larson et al. (2016) should have considered, with a sample of over 10,000 defendants, a more appropriate significance value (p) of .001. A preferable option would be to focus on effect sizes (with confidence intervals), which convey how large and meaningful a difference is, rather than merely whether it reaches "statistical significance."

We would like to explore one final thought in this section. Some readers might be wondering why anyone should care about our concerns. Discussions about ARAIs, statistics, methods, and test bias may seem complex and uninteresting (we find them rather fascinating). We are at a unique time in history. We are being presented with the chance of a generation—and perhaps a lifetime—to reform sentencing and unwind mass incarceration in a scientific way, and that opportunity is slipping away because of misinformation and misunderstanding about ARAIs. Poorly conducted research or misleading statements can lead to confusion and/or paralysis for those charged with making policy. The quote from a subsequent ProPublica article makes this point (Kirchner, 2016). Relying on the research of Angwin et al. (2016), the chair of the federal defenders legislative committee, David Patton, when being interviewed by Lauren Kirchner, posed the question "Will it be possible to validly measure those things [risk factors] for somebody who is institutionalized?" and stated "We just don't know that such a tool can be developed, or if it can, whether it will exhibit similar racial biases of current tools." In response to these issues, we analyzed a reduced set of data used by Larson et al. (2016); based on our findings, we conclude that the Larson et al. (2016) analysis was misguided and the subsequent conclusions offered by Angwin et al. (2016) are faulty. Below is a description of the methods employed to test for race-based bias with the COMPAS.

Methods

To properly test the COMPAS for race-based bias, we downloaded the dataset comprising only the sample of pretrial defendants (as the probation and parolee data were excluded) and syntax that Larson et al. (2016) used in their analyses. Two separate files were available for analysis. One file contained the information needed to test the relationship between the Northpointe COMPAS and arrest for any crime. The second file contained the information needed to test the relationship between the Northpointe COMPAS and arrest for a violent crime. We made all variable transformations in R using the same syntax as Larson et al. (2016).

We departed from their analysis in the following ways: First, we kept for analysis only those defendants whose race was either black or white. This was done as Larson et al. (2016) only mention bias between black and white defendants and doing so simplifies the analysis and subsequent discussion. This process reduced our sample sizes to 5,278 for the "any arrest" analysis file and 3,967 for the "arrest for a violent offense" analysis file with a two-year follow-up to measure recidivism.

Second, rather than analyze group mean differences to determine if bias exists, we used a framework that tests for bias in the degree of prediction as a function of race and functional form of prediction (i.e., slope and intercept) as a function of race. This framework is based on methods of testing for bias developed,

recognized, and used in other professions. The framework is reviewed and applied to a risk assessment by Skeem and Lowenkamp (2016a, 2016b) and Skeem, Monahan, and Lowenkamp (2016).

Third, the Northpointe COMPAS decile score was used in all analyses rather than the category ratings (low, medium, high). It should be noted that the results of the analyses were not dependent on the scale of the risk score. The same results were obtained when we used the decile score or the category ratings, and the decile scores provided a more refined or precise estimate than the risk categories (e.g., low, medium, high).

To test for bias in the degree of prediction as a function of race, we calculated AUC-ROC values for the overall sample and for each race.[4] The AUC-ROC values for white and black defendants were then compared using z-tests. We calculated and analyzed AUC-ROC values using any arrest as the outcome measure and then using arrest for a violent offense as the outcome measure.

To test for bias in form as a function of race, we calculated a series of logistic regression models predicting each of the outcomes (first for any arrest and then for arrest for a violent offense). To test for bias in form, we inspected interaction terms between the race and the Northpointe decile score for each outcome of interest. In addition, the magnitude and statistical significance of the coefficient for race was inspected for each outcome.

Results

Initial analyses involved an examination of general recidivism base rates for the sample and then across race. Results of these analyses are presented in Table 1, which shows the base rate of failure (general rearrest) as 47 percent for all defendants, 39 percent for White defendants, and 52 percent for Black defendants. It is important to note that the general recidivism base rate for Black defendants is significantly higher than it is for White defendants specifically, and the overall sample generally. Racial differences in failure rates across race describe the behavior of defendants and the criminal justice system, not assessment bias. Results also indicate that failure rates seem to monotonically increase with the risk categorizations of the COMPAS in that 29-35 percent of low-risk defendants were rearrested (White and Black respectively), 53-56 percent of medium-risk defendants were rearrested (respectively), and 73-75 percent of high-risk defendants were rearrested (also respectively). Note here that while the base rate of general recidivism differed significantly for White and Black arrestees (with Black defendants evidencing higher rearrest rates), the general recidivism failure rates for White and Black defendants are somewhat similar across low-, medium-, and high-risk categorizations.

To explore the predictive fairness of the COMPAS, we first examined whether the degree of the relationship between COMPAS scores and general recidivism varied due to race. Analyses of the degree of accuracy involved AUC-ROC analyses, which are appropriate for accomplishing this goal because they identify the chance (or probability) that a randomly selected arrestee will have a higher COMPAS score than will a randomly selected non-arrestee. AUC-ROC values range from zero to one, with .5 indicating mere chance prediction (or "fifty-fifty"), 1 indicating perfect prediction, and AUC-ROC values of .56, .64, and .71 signifying small, medium, and large predictive benchmark effects, respectively (Rice & Harris, 2005). As an interpretive example, an AUC-ROC value of .71 would translate to a randomly selected arrestee scoring higher on the COMPAS than would a randomly selected non-arrestee 71 percent of the time. If the COMPAS is differentially accurate in its degree of recidivism prediction across race, corresponding z-tests derived from AUC-ROC values for White and Black arrestees will be significantly different from one another. The following analyses are those that comport with accepted standards for determining if a particular test is biased against a particular group.

Degree of Relationship

In accordance with standard practices in testing for bias on education and psychological tests, the AUC-ROC values were generated and compared for the entire sample and for each group of race. AUC-ROC analyses presented in Table 1 show a moderate to strong degree of predictive accuracy for all defendants, as well as across defendant race. The COMPAS demonstrated a strong degree of accuracy in prediction for all defendants, with an AUC of .71. The AUC estimate for White defendants was .69 and .70 for Black defendants, with no significant difference between values by race. This simple lack of difference in predictive utility for the COMPAS by race contradicts the conclusions reached by Larson et al. (2016).

Table 1 also presents DIF-R values for the sample and across race to investigate the dispersion of recidivism base rates across risk categorizations of the COMPAS (as opposed to COMPAS decile score accuracy, which was examined above using AUC-ROC analyses). The values of the dispersion index for risk (or DIF-R) range from one to infinity, with larger values indicating greater accuracy, across and within each risk category as a function of base rate dispersion (Silver, Smith, & Banks, 2000). Results of the DIF-R analyses support the COMPAS risk categorizations as unique from one another and meaningful. The calculated DIF-R values in Table 1 are consistent with those found in other risk assessment studies.

TABLE 1.
Failure Rates, AUC-ROC, DIF-R for General Recidivism

	All	White	Black
Low	32	29	35
Medium	55	53	56
High	75	73	75
Base Rate*	47	39	52
AUC	0.71	0.69	0.70
DIF-R	0.73	0.65	0.70

*= $\chi^2(2) = 88.85$; $p < 0.001$

Table 2 shows the degree of prediction for the COMPAS and violent recidivism. Analyses performed were identical to those just presented above in Table 1, save for the different outcome. Failure rates for violent recidivism were 17 percent for the sample, 12 percent for White defendants, and 21 percent for Black defendants. Violent recidivism failure rates across risk categories increased with risk categorization successively, although Black defendants were arrested for a violent offense at a much higher rate than White defendants across all three categories of risk. Again, note that different (higher) violent arrest rates for Black defendants than White defendants is not an indicator of assessment bias. As noted above for general recidivism in Table 1, AUC-ROC analyses show moderate to strong and statistically similar predictive accuracy for both Black and White defendants. Further, DIF-R values for violent arrest evidence acceptable base-rate dispersion for the sample and across race, with slightly better

[4] We chose AUC-ROC as it is recognized as a standard measure in assessing diagnostic accuracy of risk assessments and has properties that make it not affected by base rate or sample size (Rice & Harris, 2005).

TABLE 2.
Failure Rates, AUC-ROC, DIF-R for Violent Recidivism

	All	White	Black
Low	11	9	13
Medium	26	22	27
High	45	38	47
Base Rate*	17	12	21
AUC	0.71	0.68	0.70
DIF-R	0.63	0.47	0.64

*= $\chi^2(2) = 49.41$; $p < 0.001$

risk category dispersion for Black defendants.

The above examination of failure rates, degree of predictive accuracy, and base rate dispersion across race fails to support the conclusions of racial bias made by Angwin et al. (2016) and, instead, finds a degree of prediction that is remarkably consistent for both Black and White defendants.

We made the argument above that Angwin et al.'s false positive/false negative analysis of the COMPAS was flawed and present a reanalysis in Tables 3 and 4. When dealing with a risk assessment that provides more than two categories, it is recommended that tests based on a 2×2 contingency table (e.g., positive predictive value, negative predictive value, false positive rate, false negative rate) be run using a specific binning strategy. That is, a decision has to be made on how to create two groups from a three (or more) category risk assessment. Singh et al. (2011) recommend first binning the low cases as the "low-risk group" and comparing them to the moderate and high-risk offenders binned together as the "high-risk group." This would be considered a "rule-in" test. The second binning process involves combining the low and moderate-risk offenders into the "low-risk group" and comparing them to the high-risk offenders (high-risk group). This would be considered a "rule-out" test.

When this process is followed, note that the false positive rates decrease substantially when binning the low and moderate risk cases together and treating them as the "low-risk" group (or the group that would be expected to succeed). In contrast, false negative rates go up in both groups. These two reversals—a decrease in false positive rates and an increase in false negative rates—might be preferred by some, as it limits the number of individuals that are identified as "high-risk." For others with a low tolerance for recidivism and victimization, the binning process where moderate and high-risk were combined to form the "high-risk" group would be preferred. Regardless, what should be taken away from these tables is the fact that when recommended practices are followed for calculating performance indicators of predictive instruments, a somewhat different pattern of results and conclusions is drawn.

Form of Relationship

To further investigate Angwin et al.'s rather serious claims of racial bias, subsequent analyses, suggested by accepted testing standards, center on the form of the relationship between recidivism and COMPAS score. More specifically, if the algorithm upon which the COMPAS is based was to perform similarly across race, then the mathematical regression slope and intercept of that relationship should also be similar across racial subgroups (Aguinis, Culpepper, & Pierce, 2010). Put more simply, we are examining the functional form (slope and intercept) of the relationship between the COMPAS and recidivism to see whether an average COMPAS decile score of x corresponds to an average arrest rate of y across race, which is the standard for examining predictive bias.

To investigate the form of the relationship between the COMPAS and recidivism across race, we estimated four logistic regression models for each of the two outcomes (general and violent recidivism) that were then compared to determine whether slope and intercept differences exist between White and Black defendants. Table 5 presents the results of these analyses, showing that Model One predicts arrest with age, gender, and

TABLE 3.
Performance Indicators Low vs. Moderate/High

White				**Black**			
		Actual				Actual	
		NR	R			NR	R
Predicted	NR	999	408	Predicted	NR	873	473
	R	282	414		R	641	1188
FN	0.50			FN	0.28		
FP	0.22			FP	0.42		
Sensitivity	0.50			Sensitivity	0.72		
Specificity	0.78			Specificity	0.58		
PPV	0.59			PPV	0.65		
NPV	0.71			NPV	0.65		

FN = False negative rate; FP = False positive rate; PPV = Positive predictive value; NPV = Negative predictive value; NR = Not recidivist; R = Recidivist

TABLE 4.
Performance Indicators Low/Moderate vs. High

White				**Black**			
		Actual				Actual	
		NR	R			NR	R
Predicted	NR	1220	660	Predicted	NR	1303	1027
	R	61	162		R	211	634
FN	0.80			FN	0.62		
FP	0.05			FP	0.14		
Sensitivity	0.20			Sensitivity	0.38		
Specificity	0.95			Specificity	0.86		
PPV	0.73			PPV	0.75		
NPV	0.65			NPV	0.56		

FN = False negative rate; FP = False positive rate; PPV = Positive predictive value; NPV = Negative predictive value; NR = Not recidivist; R = Recidivist

TABLE 5.
Logistic Regression Models Predicting Two-Year General Recidivism (N = 5278)

	Model 1	Model 2	Model 3	Model 4
Age	0.97*	0.98*	0.99*	0.99*
Female	0.58*	0.60*	0.61*	0.61*
Black	1.45*	--	1.09	1.12
NPC Decile	--	1.30*	1.30*	1.30*
NPC Decile X Black	--	--	--	0.99
Constant	2.29*	0.42*	0.40*	0.39*
Chi Square	297.68	804.42	806.13	806.19
LL	-3500.37	-3247.00	-3246.14	-3246.11
Pseudo-R^2	0.04	0.10	0.11	0.11

Note: The two dashes '- -' in the table above indicate that the variable was not included in the model.

TABLE 6.
Logistic Regression Models Predicting Two-Year Violent Recidivism (N = 3967)

	Model 1	Model 2	Model 3	Model 4
Age	0.96*	0.99	1.00	1.00
Female	0.47*	0.57*	0.57*	0.57
Black	1.57*	--	1.24	1.21
NPC Decile	--	1.32*	1.30*	1.30*
NPC Decile X Black	--	--	--	1.01
Constant	0.66	0.09*	0.08*	0.08*
Chi Square	183.53	345.34	350.49	350.52
LL	-1725.6	-1644.70	-1642.12	-1642.11
Pseudo-R^2	0.05	0.09	0.10	0.10

Note: The two dashes '- -' in the table above indicate that the variable was not included in the model.

race; Model Two predicts arrest with age, gender, and COMPAS decile score; Model Three predicts arrest with age, gender, race, and COMPAS decile score; and Model Four predicts arrest with all of the above variables, including an interaction term for race and COMPAS decile score.

Comparisons across these four models presented in Table 5 reveal two important findings relevant to an investigation of racial bias in assessment. First, an examination of Models Three and Four indicates that the addition of the interaction term between the COMPAS and race is not significant and does not improve the prediction of general recidivism for the model overall. So, the slope of the relationship between the COMPAS and general recidivism is similar for both Black and White defendants, and race does not moderate the utility of the COMPAS to predict general recidivism. Second, a comparison of Models Two and Three shows that there are no significant racial differences in the intercept (or constant) for the relationship between the COMPAS and general recidivism. Taken together, these findings suggest that there are no significant differences in the functional form of the relationship between the COMPAS and general recidivism for White and Black defendants. A given COMPAS score translates into roughly the same likelihood of recidivism, whether a defendant is Black or White.

Similar analyses were conducted for the relationship between the COMPAS and violent recidivism and these results are presented in Table 6. As above, there is no observed evidence of assessment bias in these analyses. Specifically, the relationship between race and violent recidivism becomes insignificant once the COMPAS decile score is introduced into the logistic equation. Furthermore, the interaction term between race and COMPAS decile score in Model Four is also insignificant. As above, these findings indicate no difference in the form of the relationship between the COMPAS and violent recidivism for White and Black defendants.

As a final analysis of predictive fairness by race for the COMPAS, we calculated predicted probabilities of any arrest (general recidivism) based on regression Model Four in Table 5, grouped together those probabilities for each COMPAS decile score, and then displayed the grouped probabilities across race in Figure 1. Examination of this figure shows that the slope of the relationship between the COMPAS and general recidivism does not differ by race, although Black defendants do have higher predicted (and observed) arrest rates. Similarly, we then calculated the predicted probabilities of violent arrest based on Model Four of Table 6, grouped together those probabilities for each COMPAS decile score, and then displayed the grouped probabilities across race in Figure 2. As was observed in Figure 1, the slope of the relationship between COMPAS score and violent arrest does not differ across race (again, although Black defendants have higher predicted violent arrest rates). Taken together, these two figures further support parity in the form of the relationship between the COMPAS and rearrest (general and violent).

Finally, Figures 3 and 4 visually summarize this study's findings. The bar chart in Figure 3 shows recidivism rates for any arrest by COMPAS risk category (low, medium, and high) and across race. The figure also displays a graphed line showing the percentage of Black defendants in each risk category. The graphed line shows that the percentage of Black defendants increases along with risk categorization, meaning there are more high-risk Black defendants than there are medium risk, and more medium-risk Black defendants than there are low risk. Overall, this means that Black defendants tend to score higher on the COMPAS than White defendants. Alone, this might suggest bias. However, examination of the bar chart shows that subsequent arrest rates increase along with risk categorization for both White and Black defendants and that Black defendants have higher recidivism rates than White defendants across all three

FIGURE 1.
Predicted Probability of Any Arrest by Race

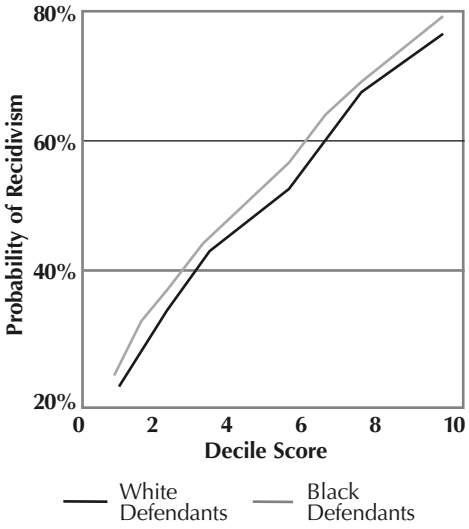


FIGURE 2.
Predicted Probability of Arrest for Violent Offense by Race

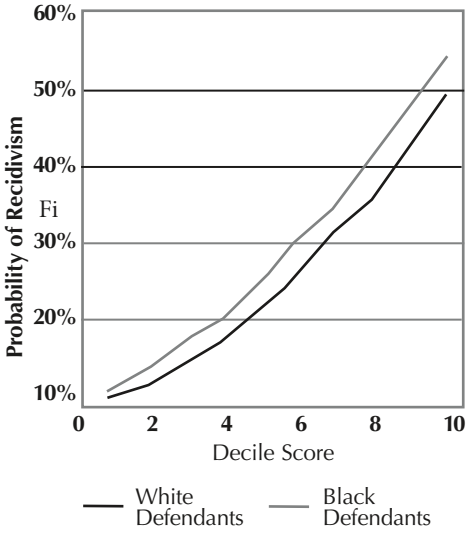


FIGURE 3.
Recidivism Rates by Race and Percent Black in Each Risk Category—Any Arrest

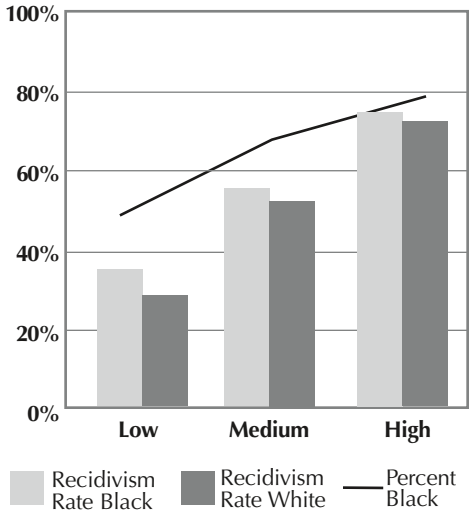
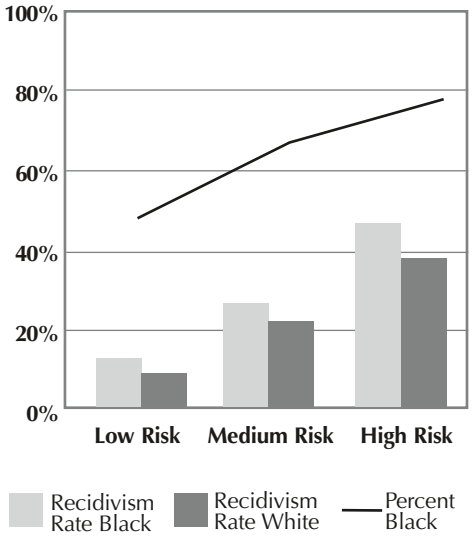


FIGURE 4.
Recidivism Rates by Race and Percent Black in Each Risk Category—Arrest Violent Offense



categories of risk.

Taken together, the two aspects of this figure show us that, despite the conclusions of Angwin et al. (2016), racial differences in mean risk scores are less indicative of test bias than of true differences in the likelihood of recidivism. The same pattern of findings also holds for violent arrest shown in Figure 4.

Discussion

A recent ProPublica.org article by Angwin et al. (2016) investigated the presence of racial bias in one of the more popular and commonly used actuarial risk assessment instruments, namely the COMPAS. The authors' conclusions are rather obvious given the title of their article: "*There's software used across the country to predict future criminals. And it's biased against Blacks.*" However, upon analyzing the same data, we came to a quite different conclusion. This section summarizes the findings of our analyses and then offers insight as to how Angwin et al. (2016) obtained different results. Ultimately, we challenge their understanding of the COMPAS and how it is to be both scored and used, their understanding of research methods and statistics, and, perhaps, their adherence to their own code of ethics.

Our initial analyses looked at the observed recidivism rates for Black and White defendants for any arrest (general recidivism) and for a violent arrest (violent recidivism). Results indicated that Black defendants were significantly more likely to be arrested for any arrest and for violent arrest. In addition, low-, medium-, and high-risk Black defendants were also rearrested more than their low-, medium-, and high-risk White defendant counterparts (for both any arrest and for violent arrest). Our second set of analyses focused on the degree of accuracy for the COMPAS in predicting any arrest and violent arrest. Our results found the COMPAS to be a good predictor of both types of arrest and, more importantly, to predict outcome equally well (i.e., of moderate degree) across both races. Furthermore, logistic regression analyses conducted to estimate the form of the relationship between the COMPAS and outcome (any arrest and violent arrest) revealed no differences in the slope and intercept, indicating that the COMPAS predicts recidivism in a very similar way for both groups of defendants. Most important, the interaction term between race and COMPAS decile score is not significant and adds no predictive power to the models overall (see Tables 5 and 6). Stated differently, the COMPAS does not predict outcome differently across groups of Black and White defendants—a given COMPAS score translates into roughly the same likelihood of recidivism, regardless of race. This may be seen visually in Figures 1 and 2. Higher mean risk scores do not indicate bias if they correspond with higher arrest rates.

In all instances, we failed to find evidence of predictive bias by race in the COMPAS. Interestingly, these findings are remarkably consistent with existing literature that has also tested for bias in other ARAIs (see Skeem & Lowenkamp, 2016a, 2016b). Had Angwin et al. (2016) conducted a thorough review of rigorous research on test bias, they undoubtedly would have discovered the existence of standards for educational and psychological testing put forth by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014). Because they failed to do so, they also failed to test for bias within these existing standards. Given the gravity of their conclusion for criminal justice policy, this failure is neither acceptable nor excusable.

However, failing to perform an exhaustive (or even cursory) literature review that might have informed their "study" is just the beginning of Angwin et al.'s (2016) shortcomings.

In addition to applying the COMPAS to an incorrect population (which in and of itself is sufficient grounds to discredit their study), Larson et al. (2016) imposed a false dichotomy on the COMPAS by reducing the risk categorizations into just two groups defined by the binary categorization of recidivist or non-recidivist. While it is problematic that they collapsed medium- and high-risk defendants into one category that was then compared against the low-risk defendants, more problematic is their interpretation of what information COMPAS scores provide (Singh, 2013). Just as medicine uses actuaries to inform patient prognoses and the auto insurance industry uses actuaries to inform probabilities of risky driving behavior, the COMPAS is based on an actuary designed to inform the probability of recidivism across its three stated risk categories. To expect the COMPAS to do otherwise would be analogous to expecting an insurance agent to make absolute determinations of who will be involved in an accident and who won't. Actuaries just don't work that way. This error discredits their main finding that Black defendants were more likely to be incorrectly identified as recidivists (false positives) while White defendants were more likely to be misclassified as non-recidivists (false negatives). Furthermore, our reanalysis of false positives and false negatives also calls into question the validity of their conclusions regarding this method of analysis when an assessment tool comprises more than just two categories (see Tables 3 and 4).

Another of their main conclusions stems from a Cox regression model predicting general recidivism with a number of variables, including an interaction term between race and COMPAS score. In this analysis, they observed a p value of .0578 for the interaction term and then concluded that race moderated the relationship between outcome and COMPAS score. This erroneous conclusion further demonstrates the carelessness in their approach, as .0578 is less than .05—particularly with a sample size of 10,000—only in the world of "data torturing" (see Mills, 1993), where authors are outright looking for something of significance to make their point.

An additional statistical oddity of Larson et al. (2016) concerns the ordering of variables in their general recidivism logistic regression model, in which they predict the COMPAS score with recidivism and a number of other demographic variables. Because assessment scores occur before recidivism, it appears as though they have their independent and dependent variables confused. We're not sure of the logic behind predicting an assessment score with recidivism but we do believe that this analysis is responsible for their conclusion that, somehow, higher average COMPAS scores for Black defendants indicate bias. Given the higher observed recidivism rates for Black defendants, and given the demonstrated validity of the COMPAS, it is nothing short of logical that these defendants evidence higher COMPAS scores (after all, isn't that precisely what the COMPAS is measuring?).

In summary, this research sought to reanalyze the study by Larson et al. (2016), using accepted methods to assess the presence of test bias. Using these accepted methods, we found no evidence of racial bias. Our analysis of Larson et al.'s (2016) data yielded no evidence of racial bias in the COMPAS' prediction of recidivism—in keeping with results for other risk assessment instruments (Skeem & Lowenkamp, in press; 2016a).

We would be remiss if we failed to report the limitations of our re-analysis of the ProPublica analysis. First, we did not completely replicate the ProPublica study, as we excluded those defendants whose race was not white or black. We also did not estimate the survival analysis models. Second, the outcome measure is limited to new arrest. The limitations (as well as strengths) for this measure have been well documented (see Maltz, 1984). Third, the extent to which the findings of this study are generalizable to other samples, jurisdictions, and other instruments is unknown. Finally, while this article was sent out to numerous colleagues for review and input, it was not a blind review and this research is yet to be published in a peer-reviewed journal.

Conclusion

It is noteworthy that the ProPublica code of ethics advises investigative journalists that "when in doubt, ask" numerous times. We feel that Larson et al.'s (2016) omissions and mistakes could have been avoided had they just asked. Perhaps they might have even asked…a criminologist? We certainly respect the mission of ProPublica, which is to "practice and promote investigative journalism in the public interest." However, we also feel that the journalists at ProPublica strayed from their own code of ethics in that they did not present the facts accurately, their presentation of the existing literature was incomplete, and they failed to "ask." We believe the result demonstrates that they are better equipped to report the research news, rather than to make the research news.

We hope that this rejoinder and its consistency with the existing literature provides some comfort (in the form of evidence) to policymakers who have been exposed to misleading information about the reliability, validity, and fairness of ARAIs. At the very least, this article highlights an accepted and legitimate approach that agencies and jurisdictions can use to determine if the ARAI they use, or are considering using, is in fact subject to predictive bias towards a particular group of people. Clearly, ARAIs hold considerable promise for criminal justice reform in that they are capable of better informing what were previously subjective and indefensible criminal justice decisions (Andrews, Bonta, & Wormith, 2006).

References

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95,* 648-680.

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.

Andrews, D., Bonta, J., & Wormith, S. (2006). The recent past and near future of risk and/or need assessment. *Crime and Delinquency, 52(1),* 7-27.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Brennan, T., Dietrich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment. *Criminal Justice and Behavior, 36(1),* 21-40.

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

Dawid, A. P. (2015). Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application, 2,* 273-303.

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.

Hanson, R. K., & Morton-Bourgon, K. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1-21.

Harris, G. T., Lowenkamp, C. T., Hilton, N. Z. (2015). Estimate for risk assessment precision: Implications for individual risk communication. *Behavioral Sciences and the Law, 33,* 111-127.

Henrichson, C., & Delaney, R. (2012). The price of prisons: What incarceration costs the taxpayer. Vera Institute of Justice. Retrieved from: http://www.vera.org/pubs/special/price-prisons-what-incarceration-costs-taxpayers

Imrey, P. B., & Dawid, A. P. (2015). A commentary on statistical assessment of violence recidivism risk. *Statistics and Public Policy, 2*(1), 1-18.

Kirchner, L. (2016). The Senate's popular sentencing reform bill would sort prisoners by risk score. Retrieved from: https://www.propublica.org/article/senates-popular-sentencing-reform-bill-would-sort-prisoners-by-risk-score

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Retrieved from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Marco, C.A. & Larkin, G. L. (2000). Research ethics: Ethical issues for data reporting and the quest for authenticity. *Academic Emergency Medicine, 7*(6), 691-694.

Maltz, M. D. (1984). *Recidivism*. Originally published by Academic Press, Inc., Orlando, FL.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Mills, J. L. (1993). Data torturing. *The New England Journal of Medicine, 329*(16), 1196-1199.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 15-20.

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: A comparison of methods. *Criminal Justice and Behavior*, *27*, 733-764.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Criminal Justice and Behavior*, *31(1),* 8-22.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violent risk assessment tools: A systematic metaregression analysis of 68 studies involving 25,890 participants. *Clinical Psychology Review, 31(3),* 499-513.

Skeem, J. L., & Lowenkamp, C. T. (2016a). Risk, race, and recidivism: Predictive bias and disparate impact. Manuscript under review. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339

Skeem, J., & Lowenkamp, C. T. (2016b). Race and actuarial risk assessment: The Level of Service Inventory Revised. University of California, Berkeley: Unpublished Manuscript.

Skeem, J. L., Monahan, J., & Lowenkamp, C. T. (in press). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior*

United States Census Bureau (2014). United States Quick Facts. Retrieved from: https://www.census.gov/quickfacts/table/PST045215/00

Wagner, P., & Walsh, A. (2016). States of incarceration: The global context. Prison Policy Initiative. Retrieved from http://www.prisonpolicy.org/global/2016.html