Accuracy and Racial Biases of Recidivism Prediction Instruments

Julia J. Dressel

Senior Honors Thesis

Advisor: Professor Hany Farid

Dartmouth Computer Science Technical Report TR2017-822

May 31, 2017

Abstract

Algorithms have recently become prevalent in the criminal justice system. Tools known as recidivism prediction instruments (RPIs) are being used all over the country to assess the likelihood that a criminal defendant will reoffend at some point in the future. In June of 2016, researchers at ProPublica published an analysis claiming an RPI called COMPAS was biased against black defendants. This claim sparked a nation-wide debate as to how fairness of an algorithm should be measured, and exposed the many ways that algorithms are not necessarily fair. Algorithms are used in the criminal justice system because they are regarded as more accurate and less biased than human predictions; however, there does not exist a contemporary comparison of the performance of human and algorithmic recidivism predictions. To address this, we set out to determine if COMPAS is more accurate than human prediction, and to identify how the racial biases of human recidivism predictions compare to the racial biases of the COMPAS algorithm. After establishing a baseline performance of human prediction, we explore whether incorporating human judgment into algorithms can enhance prediction accuracy.

Contents

1. Introduction

With the recent rise of Big Data and the prevalence of technology in our everyday lives, humans have become frequent subjects of algorithms. Data is collected on all of us constantly, and many of our daily experiences are determined by algorithmic systems. Some of these algorithms affect our daily interactions with technology. Spotify uses algorithms to make personalized song recommendations [6]. Google builds algorithms to determine the ads that each user sees [9]. Other algorithms affect an individual's future education or employment. An education consulting firm called Noel-Levitz provides an analytical tool called Forecast Plus that ranks prospective students applying to a university [31]. Workforce Ready HR screens job candidates using personality tests to predict which individuals will perform better and which will remain with the company for longer [31]. Many employers use automated systems to sort through résumés and to determine optimally efficient work schedules for their employees [31]. Banks and online lenders have started making loan decisions using algorithmic predictions [12]. In Washington D.C., public schools rely on algorithmic assessment tools to identify and fire low-performing teachers [31].

Recently, algorithms have been increasingly used in the criminal justice system. The police department of Reading, Pennsylvania, for example, uses a crime prediction software called PredPol to predict where crime will most likely occur in the city during each hour of the day [31]. This practice, known

as predictive policing, occurs in various cities all over the country. In Chicago, the police department developed technology to identify the top four hundred people most likely to commit a violent crime in the city, ranked by the probability that they would be involved in a homicide [31]. There was even software developed in China that claims it can predict whether someone is a criminal based on their facial features alone [33].

Certain types of tools known as risk assessment instruments have become particularly prevalent in the criminal justice system. Some instruments predict the likelihood that an individual will fail to appear at their court hearing [1]. Recidivism prediction instruments (RPI's) assess the likelihood that a defendant will recidivate, defined to mean that they will reoffend at some point in the future. These tools use an individual's criminal history, personal background, and demographic information to make these risk predictions. Various risk assessment instruments are being used across the country to inform pretrial decisions, parole decisions, and sometimes even sentencing decisions [10]. The extensive use of this technology, however, is not without controversy.

One of the criminal risk assessment tools used across the country is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [5]. Built by the for-profit company Northpointe Inc.[1], COMPAS has assessed over 1 million offenders since it was developed in 1998 [30]. The RPI component of COMPAS is called the Recidivism Risk Scale, which has been used in COMPAS since 2000 [7]. This scale is computed from 137 questions,

[1]In January of 2017, Northpointe Inc. rebranded to the name "equivant" [2]. Since the debate surrounding the COMPAS algorithm has used the name Northpointe Inc., we will refer to the company as Northpointe in this paper.

which are either personally answered by a defendant or are determined automatically from a defendant's criminal record [5]. According to the COMPAS practitioner's guide, this recidivism risk scale is built to predict "a new misdemeanor or felony offense within two years of the COMPAS administration date" [30].

In May of 2016, Angwin et al. at ProPublica released an analysis [5] of the COMPAS Recidivism Risk Scale. The goal of the analysis was to assess the accuracy of the COMPAS recidivism prediction algorithm, and to investigate whether the tool was biased against any particular group. Angwin et al. gathered data on over 7,000 individuals arrested in Broward County, Florida during 2013 and 2014, and determined whether each individual was charged with a new crime over the next two years [5]. Because Broward County uses the COMPAS recidivism risk scores to determine whether a defendant should be released or detained before their trail, ProPublica collected the COMPAS scores given to each defendant before their trial [25]. Using this data, ProPublica was able to assess the accuracy and bias of the COMPAS recidivism risk scores.

COMPAS scores range from 1 to 10. Scores of 1-4 are labeled Low-Risk, scores of 5-7 are labeled Medium-Risk, and scores of 8-10 are labeled High-Risk. ProPublica considered scores higher than "Low-Risk" to indicate a risk of recidivism [25], citing the Northpointe practitioner's guide to COMPAS that states, "scores in the medium and high range garner more interest from supervision agencies than low scores, as a low score would suggest there is little risk of general recidivism" [30]. Therefore, ProPublica translated the COMPAS

	Individual Did Recidivate	**Individual Did Not Recidivate**
COMPAS Predicticed Will Recidivate	True Positive	False Positive
COMPAS Predicted Will Not Recidivate	False Negative	True Negative

Table 1.1: Classification in the COMPAS binary prediction scheme.

scores into a binary prediction that a defendant would recidivate or would not recidivate. In binary prediction circumstances, errors are classified as either false positive errors or false negative errors. In this binary prediction scheme, the positive prediction is that an individual will recidivate. The negative prediction is that an individual will not recidivate. Therefore, a false positive error occurs when an individual is classified by the algorithm as someone who will recidivate, but they do not actually recidivate. A false negative error occurs when an individual is predicted to not recidivate, but they do recidivate, Table 1.1.

ProPublica was specifically interested in how the COMPAS algorithm treated black defendants as compared to white defendants, which are the two largest racial demographics in their data set[2]. They compared the accuracy of COMPAS on the two groups, which is defined as the percentage of individuals that COMPAS correctly classified as a recidivator or non-recidivator. They found that COMPAS was similarly predictive for both races (66.9% accurate for white defendants, and 63.8% accurate for black defendants) [25]. The algorithm, however, made drastically different mistakes on each racial group. Black defendants who did not recidivate were almost twice as likely to be classified as recidivators compared to white defendants who did not recidivate.

[2]51.2% of the defendants in the data set are black. 34.0% are white.

Defendant Race	Accuracy (%)	False Positive Rate (%)	False Negative Rate (%)
Black	63.8	44.9	28.1
White	66.9	23.5	47.7

Table 1.2: COMPAS performance by race.

Of the black defendants who did not go on to recidivate, 44.9% of them were misclassified as recidivators. Of the white defendants who did not go on to recidivate, only 23.5% of them were misclassified as recidivators. The opposite is true for defendants who did go on to recidivate. Of the black defendants who did go on to commit a new crime, 28.1% of them were misclassified as low-risk. Of the white defendants who did commit a new crime, 47.7% of them were misclassified as low risk. This means that white individuals who did recidivate were almost twice as likely to be given a low risk score compared to black individuals who did recidivate. In other words, ProPublica found that the false positive error rate for black defendants was almost twice as high as the false positive error rate for white defendants. Conversely, the false negative error rate for white defendants was almost twice as high as the false negative error rate for black defendants.

In this scenario, a false positive error is detrimental to a defendant, whereas a false negative error is beneficial for a defendant. Being misclassified as high-risk may result in a defendant being detained in jail before their trail, given a higher bail, or even given a longer sentence. Being misclassified as low-risk means that an individual may suffer less harsh consequences. Thus, the imbalanced false positive and false negative error rates of the COMPAS algorithm are benefitting white defendants and unfairly punishing black defen-

dants. This situation is known as *disparate impact*, where a penalty system has "unintended disproportionate adverse impact on a particular group" [10]. Because black defendants were suffering from the COMPAS errors and white defendants were benefitting from the COMPAS errors, ProPublica reported that the COMPAS algorithm was biased against blacks.

In July of 2016, less than two months after ProPublica published their critique, Northpointe responded with a 39-page rebuttal [15] to ProPublica's claims. Primarily, Northpointe argues that the ProPublica authors overlooked the fact that the COMPAS risk scales satisfy *predictive parity*. An algorithm satisfies predictive parity if it generates equally accurate predictions for all groups [4]. Because COMPAS was around 60% accurate for both whites and blacks, Northpointe argues that it was fair to both groups. This argument rests in the belief that a recidivism prediction instrument should not be more accurate for one group than it is for another. Furthermore, Northpointe defends that COMPAS satisfies *accuracy equity*, which means that it can "discriminate recidivists and non-recidivists equally well for two different groups" [15]. The most commonly used measure of discriminative performance is the area under the receiver operating characteristic (ROC) curve. This is known as an AUC or AUC-ROC value, and, in this context, represents "the probability that a randomly selected recidivist will have a higher risk score than a randomly selected non-recidivist" [15]. AUC values of 0.70 or above are considered to indicate satisfactory discriminative ability [7]. Northpointe reports in their rebuttal that the AUC value for white defendants is 0.693, and the AUC value for black defendants is 0.704. This suggests that COMPAS accomplishes satis-

factory predictive accuracy on both white and black defendants. Additionally, Northpointe claims that the difference in these AUC values is not significant (p=0.483). Therefore, Northpointe maintains that it can discriminate recidivists from non-recidivists equally well for both black and white defendants. Finally, Northpointe claims that equal false positive and false negative rates are both an "unrealistic criterion" [15] as an assessment of racial bias, because the two groups have different base rates. In other words, because a greater percentage of black defendants in the data set recidivated compared to white defendants (52.3% compared to 39.1%), it is unreasonable to expect the false positive and false negative rates to be the same for both groups.

These contradictory notions of fairness sparked a nation-wide debate as to how fairness should be measured [11, 26, 35, 3, 10, 24, 20, 17]. One paper published in September in the journal *Federal Probation* defended the use of AUC values to prove non-discrimination, citing that AUC values are the standard measure of risk prediction performance [17]. This paper accused ProPublica of not using the "accepted methods to assess the presence of test bias" [17]. In October, researchers at Google proposed two criteria for measuring unfairness: *equalized odds* and *equalized opportunity* [20]. A prediction algorithm satisfies equalized odds if it has equal true positive rates and equal false positive rates for each group of individuals. This criterion aligns with ProPublica's notion of fairness, that the types of errors made in the predictions must be equally frequent for both white and black defendants. The requirement of equalized opportunity assumes there is a desired outcome of the prediction algorithm, such as being granted a loan or being accepted into a school. A binary pre-

dictor satisfies equalized opportunity if the probability of receiving the desired outcome (assuming one is qualified for that outcome) is independent of one's group membership. In the case of recidivism predictions, the desired outcome for an individual is being rated as low-risk. Therefore, a recidivism predictor satisfies equalized opportunity if the true negative rates are equal for both white and black defendants. Equalized opportunity is a weaker measurement of fairness, and may not apply well to the case of recidivism predictions, because the undesired outcome (receiving a high-risk score) can have significant negative consequences.

Kleinberg et al. at Cornell University [24] and Chouldechova at Carnegie Mellon University [10] independently released papers in September and October, respectively, arguing that the COMPAS risk scores are *well-calibrated*, which is a crucial requirement for a fair algorithm. For risk scores to be well-calibrated, each score must have the same meaning regardless of a defendant's race [24]. In other words, white and black defendants with the same risk score should be equally likely to reoffend (see Figure 1.1). This is an important requirement of risk scores, because judges should not have to consider a defendant's race when interpreting their score.

Both Kleinbert et al. and Chouldechova discussed why balancing false positive and false negative rates across groups (or, achieving equalized odds) is also a desired property of fairness. If these error rates are significantly different between groups, then the predictions negatively impact one group more than the other. However, Kleinbert et al. and Chouldechova both arrived at an intriguing conclusion: unless two groups have equal base rates, building

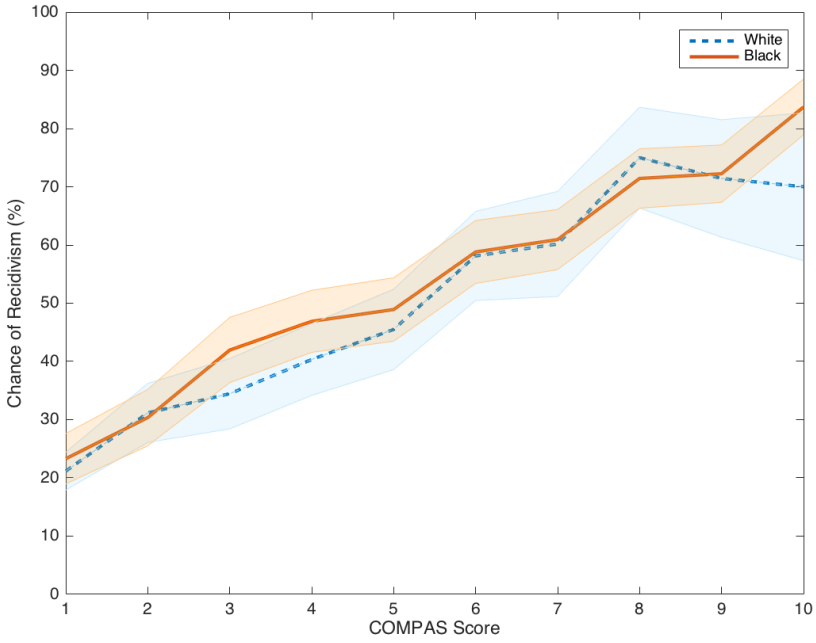

Figure 1.1: Recidivism rate by COMPAS risk score, for white and black defendants. The shaded areas represent 95% confidence intervals at each value.

an algorithm that produces well-calibrated scores as well as equal false positive and false negative rates across groups is mathematically impossible[3]. In the context of recidivism prediction, this means that algorithms that predict recidivism cannot be both well-calibrated and satisfy equalized odds unless

[3]It is also possible to satisfy these three fairness conditions simultaneously if the algorithm achieves perfect prediction, but this is an unreasonable expectation.

the recidivism rate among white individuals is the same as that among black individuals.[4]

Clearly, the debate sparked by ProPublica's article exposed the lack of a cohesive statistical definition for fairness. As Chouldechova explains, "it is important to bear in mind that fairness itself is a social and ethical concept, not a statistical one" [10]. All of the fairness criteria discussed have reasonable ethical grounds to be measures of fairness. However, as Kleinbert et al. and Chouldechova revealed, some definitions of fairness are mathematically incompatible. Thus, defining a cohesive definition of fairness that is mathematically plausible may be futile.

Algorithmic predictions have become common in the criminal justice system because they maintain a reputation of being objective and unbiased, while human decision making is considered inherently biased and flawed. Northpointe describes COMPAS as "an objective method of estimating the likelihood of reoffending" [30]. The Public Safety Assessment (PSA), a pretrial risk assessment tool developed by the Laura and John Arnold Foundation, advertises itself as a tool to "provide judges with objective, data-driven, consistent information that can inform the decisions they make" [22]. In general, people often assume that algorithms using "big data techniques" are unbiased simply because of the amount of data used to build them [29]. However, as the debate on fairness exposed, algorithms are not necessarily fair. Thus, it may be

[4]Within the current criminal justice circumstance, it may be unreasonable to expect equal recidivism rates in white and black populations in the United States. The U.S. Department of Justice published a report in June of 2016 stating that 39.7% of white prisoners released from federal prisons in 2005 were arrested for a new crime by 2010, compared to 55.1% for black prisoners. 73.1% of white prisoners released from state prisons in 2005 were arrested for a new crime in 2010, compared to 80.6% for black prisoners [27].

dangerous to assume that an algorithm is fundamentally more objective than human judgment, for this assumption could camouflage bias under the guise of technology.

Algorithmic predictions are also regarded as inherently more accurate than human predictions. The literature surrounding risk assessments have accepted that risk assessment instruments are definitively more accurate than human judgment [10, 17]. Many studies that have compared algorithmic risk assessments to professional judgment have found that data-driven risk assessments are usually more accurate the human predictions [18]. However, most of the studies that have compared algorithmic and human prediction have focused on medical diagnoses circumstances. The most recent study that explored recidivism prediction was done in Canada in 1984 [36]. As of this writing, we are not aware of a contemporary comparison of the accuracies of human versus algorithmic recidivism predictions.

Algorithmic recidivism predictions are increasingly utilized in the criminal justice system because they are considered less biased and more accurate than human predictions. As we have seen, algorithms are not inherently objective. Consequently, if algorithmic predictions are used in the criminal justice system, they should be at least as accurate as human judgment, otherwise they maintain no advantage over human predictions. Thus, it is imperative that humans are tested to evaluate their performance in predicting recidivism.

This study set out to determine if COMPAS is more accurate than human prediction. Additionally, we will identify how the biases of human predictions compare to the biases of the COMPAS algorithm. Understanding the baseline

accuracy and biases of human recidivism predictions will provide a standard above which algorithmic predictions should perform in order to justify their continued use in the criminal justice system.

2. Human Recidivism Prediction

The primary goal of this study is to determine how accurately humans can predict recidivism compared to the COMPAS algorithm. Additionally, this study aims to compare the racial biases of human recidivism predictions with COMPAS predictions.

2.1 Methods

We downloaded the data set provided by ProPublica in their analysis of COMPAS. This data set contains information on 7,214 pre-trial defendants in Broward County, Florida during 2013 and 2014. The data set contains each defendant's demographic information, criminal history, the crime for which they were arrested, their COMPAS "Risk of Recidivism" score, and whether they were arrested for a new crime within two years of their COMPAS screening.

COMPAS scores range from 1 to 10. Scores of 1-4 are labeled Low-Risk, scores of 5-7 are labeled Medium-Risk, and scores of 8-10 are labeled High-Risk. Therefore, in this study, a COMPAS "Risk of Recidivism" score of 5 or

higher is considered a prediction that the individual would recidivate. This is the same cutoff used by ProPublica.

Of the 7,214 defendants in the data set, 1,000 were selected for use in this study. A total of 1,000 defendants were randomly selected from the data set until the COMPAS accuracies, false positive rates, and false negative rates on the subset were representative of COMPAS' performance on the entire data set. A descriptive paragraph about each defendant was generated using the following data fields:

1. Sex
2. Age
3. Race
4. Criminal Charge Description
5. Criminal Charge Degree
6. Juvenile Misdemeanor Count
7. Juvenile Felony Count
8. Number of Prior Crimes

Each description paragraph followed this format:

> The defendant is a [RACE] [SEX] aged [AGE]. They have been charged with: [CRIME_CHARGE_DESCRIPTION]. This crime is classified as a [CRIME_DEGREE]. They have been convicted of [PRIORS_COUNT] prior crimes. They have [JUV_FEL_COUNT]

juvenile felony charges and [JUV_MISD_COUNT] juvenile misdemeanor charges on their record.

In this subset of 1,000 defendants, there were 63 unique criminal charge descriptions (see Appendix A). We gathered more detailed descriptions of each of these crimes according to Florida laws [23]. The participants making recidivism predictions are assumed to have no background in criminology, so a brief explanation of the crime was given to help with their predictions. These short crime descriptions followed this format:

[CRIME NAME]: [SHORT CRIME DESCRIPTION]

The goal of the study is to measure how accurately a person can predict a defendant's future recidivism based on only this brief information of a defendant. For each defendant, a participant was shown the description paragraph about the individual, as well as the short explanation of the crime with which the individual was charged. The participant was then asked, "Do you think this person will commit another crime within 2 years?" The participant responded by selecting either "Yes" or "No". The participants were required to answer every question, and they could not change a response once it was submitted.

The 1,000 defendants were randomly divided into 20 blocks of 50 defendants. Each participant was randomly assigned to see 1 of these 20 blocks. Therefore, each participant made predictions on 50 different defendants. The participants saw the 50 questions in a random order.

After making a prediction, the participant was told if their answer was correct or incorrect. After each response, the participant was shown their overall accuracy.

This study was run through Amazon's Mechanical Turk, an online crowdsourcing marketplace where individuals are paid to perform short tasks. Our task was titled "Predicting Crime" with the description: "Read a few sentences about an actual person and predict if they will commit a crime in the future". The keywords for the task were "survey, research, criminal justice". We offered $1 to each participant who completed our task. When a Turk participant decided to complete our task, they were directed to a survey, which was powered by Qualtrics. For the remainder of the paper, the term "task" will be used to describe the survey that the participants took.

At the start of the task, the participants were given this brief explanation of the study:

> In this exercise, you will be shown information regarding an individual who has previously committed a crime but is not currently incarcerated. You will see this person's age, race, gender, previous criminal history, and a description of the crime with which they were most recently charged. After evaluating this information, you will predict whether this person will commit another crime within 2 years.
>
> The information you will see is from real people and cases. You will see a random selection of 50 people from a large database of over 5,000 entries.

> We are performing this study to determine how accurately humans can determine the risk that a defendant poses. **Your performance on this task may very well inform our Courts and so it is important that you try your best.**
>
> You will be paid $1 for completing this task. If your overall **accuracy is greater than 65% we will pay you a $5 bonus** for a total of $6 for completing this task.

The $5 bonus was intended to provide an incentive to the participants to pay close attention to the task. We chose the cutoff value to be 65% because the accuracy of COMPAS on the 1,000 defendants included in the study is 65.2%. Additionally, we told the participants that their performance could inform the Courts, which provided a moral incentive for the participants to pay attention to the task.

Before beginning, the participant was also shown a sample question (Figure 2.1) to prepare them for the task. They were also shown an example of what the feedback to each question would look like (Figure 2.2).

Once the participant read the task instructions, they were asked to provide their age, gender, race/ethnicity, and education level, Appendix B. After reading the instructions and answering the demographic questions, the participant was presented with 50 questions, in a random order. The participant saw only one question at a time, and was given feedback on their answer after each question. The participant's current accuracy was displayed at the top of the screen, Figure 2.3.

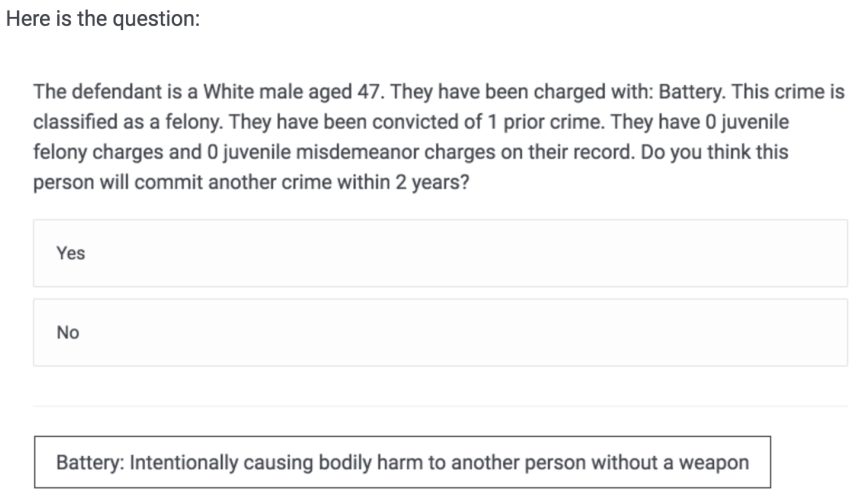

Figure 2.1: Sample question shown to participants.

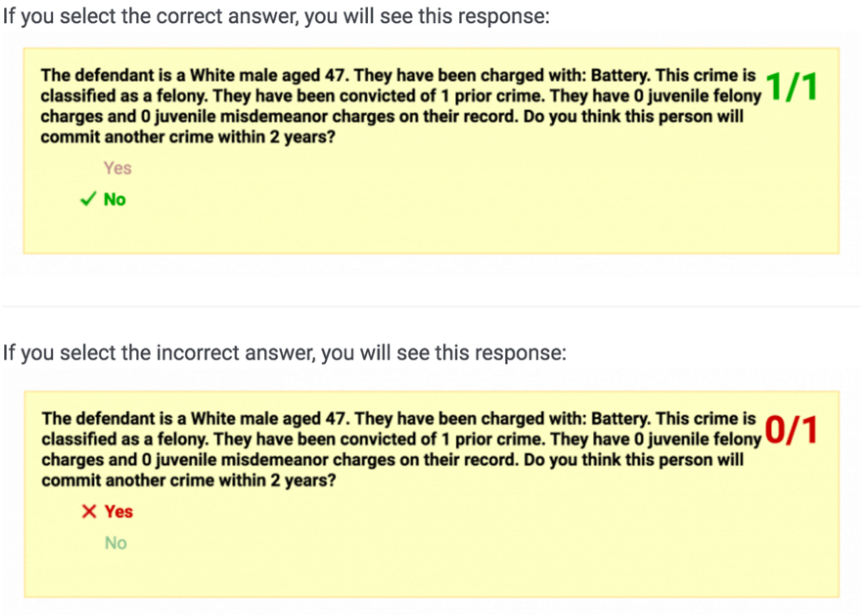

Figure 2.2: Feedback shown to a participant after each question in the task.

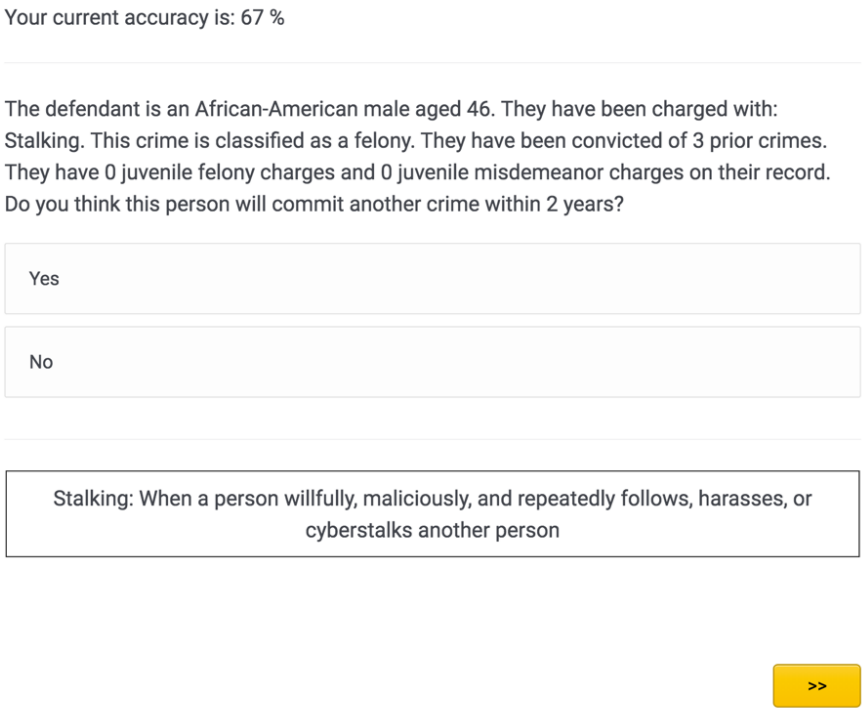

Figure 2.3: The participant's current accuracy was displayed above each quesion.

The participants were shown 3 additional catch-trial questions at random points during the task to check if they were paying attention. These questions were formatted to look similarly to the other questions, but had easily identifiable correct answers, Appendix C. If a participant answered any of these questions incorrectly, their responses were not counted in our study. The answers to these questions contributed to the overall accuracy that the participant was shown. However, these questions were not counted in our analysis of the participants' accuracies.

At the end of the task, participants were shown their final accuracy. Finally, the participants were given a unique 8-digit code to submit to Mechanical Turk. These codes were used to associate a Mechanical Turk participant with their task submission.

A total of 457 responses were recorded. All participants were located in the United States when they completed the task. There were 49 responses removed from the results due to incorrect answers on the catch-trial questions.

Responses were collected until each block of 50 questions had been seen by 20 different valid observers. After removing the catch-trial failures, 408 responses remained. For the blocks that had more than 20 valid responses, a random 20 responses were kept in the results. Ultimately, there were 400 responses. Each block of 50 defendants was seen by 20 participants.

2.2 Results

The 400 participants correctly predicted criminal recidivism with a median accuracy of 64.0%, and a mean accuracy of 62.3%. The distribution of the 400 participant accuracies is shown in Figure 2.4.

The accuracy of the participants was analyzed in comparison to the COMPAS performance on the same subset of defendants. Because groups of 20 participants were shown the same block of 50 defendants, the individual participant accuracies are not independent from one another. However, the median accuracies on each block of 50 defendants can reasonably be assumed to be independent. Therefore, the median participant performance on the 20

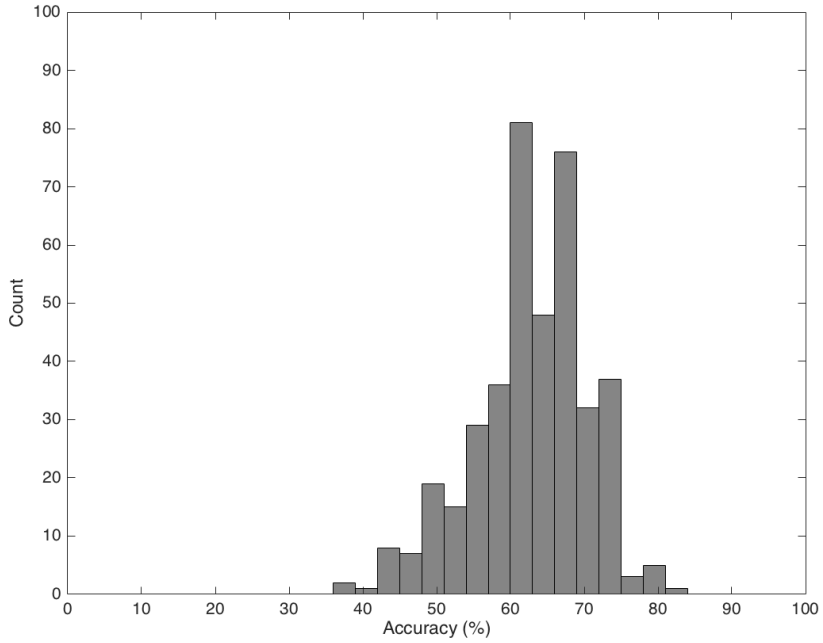

Figure 2.4: Participant accuracy distribution.

different blocks can be directly compared to the COMPAS performance on the same 20 blocks. Shown in Table 2.1 is the median participant accuracy on the 20 blocks of 50 defendants in the study, as well as the accuracy of COMPAS on each of the 20 blocks. The participant results are subject to low outliers, because participants may have given up on the task once they realized they were no longer eligible for the bonus. This behavior skews the mean accuracies to be lower, so the median accuracies of participants were considered instead of the mean accuracies. As previously noted, a COMPAS score greater than 4 was considered a prediction that an individual would recidivate, and a COM-

Block Number	Median Participant Accuracy (%)	COMPAS Accuracy (%)
1	56	60
2	64	64
3	52	60
4	69	76
5	63	68
6	67	74
7	70	66
8	67	66
9	67	68
10	63	66
11	60	70
12	60	62
13	60	62
14	68	62
15	62	60
16	63	66
17	60	62
18	66	52
19	64	70
20	60	70

Table 2.1: Participant and COMPAS accuracy per block of 50 defendants.

PAS score of 4 or less was considered a prediction that the individual would not recidivate.

Across these 20 blocks, the average of the 20 median participant accuracies was 63.1%, while the mean COMPAS accuracy was 65.2%. A matched-pairs t-test was performed to determine if the difference was significant. The test statistic was not significant at the 0.05 critical alpha level, $t(19)=-1.6690$, $p=0.1115$. Therefore, there is not sufficient evidence to suggest that COMPAS is significantly more accurate than the participants.

The concept of "the wisdom of crowds" is that the judgments of a crowd of individuals can outperform individuals that are not experts in what they are judging [34]. The predictive power of the crowd has been demonstrated in multiple studies [34, 21]. The participant responses of this study were analyzed to determine if the collective responses of all observers are more accurate than

the individual responses. These collective responses will be referred to as the "crowd".

Each of the 1,000 defendants was seen by a group of 20 participants. Each of these participants made a "Yes" or "No" prediction on the defendant. Thus, we can examine the percentage of participants who answered "Yes" on a defendant to determine the prediction accuracy of the crowd. If more than 50% of participants answered "Yes" on a defendant, that was considered a prediction that the individual would recidivate. Conversely, if more than 50% of the participants guessed "No" on a defendant, that was considered a prediction that the defendant would not recidivate. A cutoff of 50% follows the "majority rules" group prediction strategy that has been shown to be both efficient and robust [21].

Using this mode of prediction, the crowd achieved an accuracy of 66.5% across the 1,000 defendants, slightly higher than the median individual response of 64.0%, and the COMPAS accuracy of 65.2%. The crowd performance was analyzed in comparison to the COMPAS performance on the 1,000 defendants. A matched-pairs t-test was performed to determine if the difference was significant. Again, the accuracies were compared according to the accuracy on each of the 20 blocks of 50 defendants. The test statistic was not significant at the 0.05 critical alpha level, $t(19)=1.0467$, $p=0.3083$. Therefore, there is not sufficient evidence to suggest that the difference between the crowd accuracy and the COMPAS accuracy is statistically significant.

The crowd performance can also be assessed according to its AUC-ROC value. As previously discussed, the AUC-ROC (area under the receiver operat-

ing characteristic curve) is the most commonly used measure of discriminative performance for risk assessments. AUC-ROC is widely used because it is not affected by base rates or sample sizes [17]. The AUC-ROC value of a recidivism classifier represents "the probability that a randomly selected recidivist will have a higher risk score than a randomly selected non-recidivist" [15]. In this context, the percentage of the crowd that guessed "Yes" on a defendant represents the defendant's "risk score". Thus, the AUC-ROC value of the crowd predictions represents the probability that a randomly selected recidivist will have a higher percentage of "Yes" votes from the crowd than a randomly selected non-recidivist.

Shown in Figure 2.5 are the ROC curves of both the crowd and COMPAS on the 1,000 defendants. The AUC-ROC value of the crowd was 0.7101 ± 0.03, and the AUC-ROC value of COMPAS was 0.7043 ± 0.035. AUC values of 0.70 are considered to indicate satisfactory predictive accuracy [7], so both the crowd and COMPAS performed satisfactorily on this set of defendants. The most common test for comparing AUC-ROC values is the method from DeLong et al. (1988) [13]. These two AUC-ROC values were compared using this method, and the difference between the values was not significant at the 0.05 critical alpha level, $p=0.7257$. Therefore, there is not sufficient evidence to suggest that the difference between the crowd and COMPAS AUC-ROC values is significant.

The predictive accuracies of various observer demographics were analyzed to determine whether one's demographic affects their accuracy in predicting

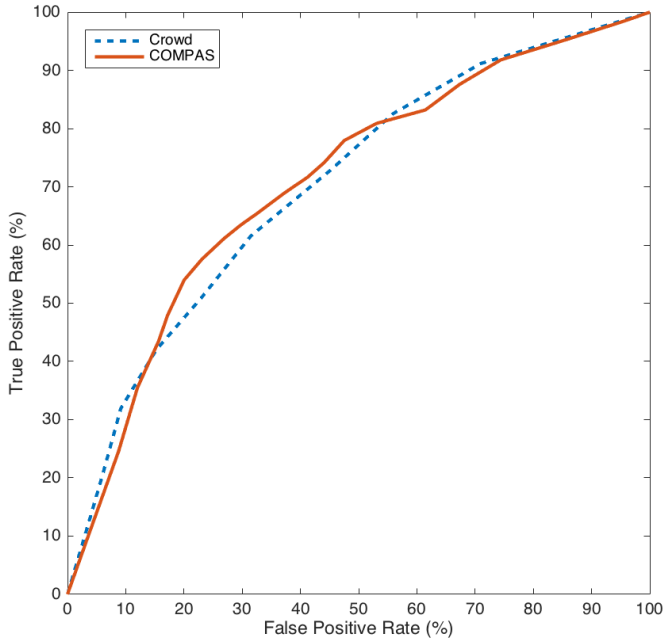

Figure 2.5: Receiver Operating Characteristic Curves of COMPAS and the crowd.

recidivism. We found that one's demographic did not affect their accuracy in predicting recidivism. Details of this analysis can be found in Appendix D.

2.2.1 Fairness

Although there does not exist a standard measurement of fairness, we can evaluate the fairness of the participants in terms of how well their results meet the following fairness criteria: Predictive Parity; Equalized Odds; Accuracy Equity; and Calibration.

Predictive Parity An algorithm satisfies *predictive parity* if it is equally accurate across all groups [4]. Therefore, the Mechanical Turk participants satisfy predictive parity if their responses were equally accurate for white and black defendants. Across the 400 individual participant responses, the average accuracy on black defendants was 62.3% (median = 63.0%), and the average accuracy on white defendants was 63.5% (median = 65.0%). Because the individual participant accuracies aren't independent, they must be analyzed by independent block accuracies. Therefore, one can compare the accuracies of the participants on black defendants in each block to the accuracies on white defendants in each block. Again, the median accuracies for each block were used. The average of the median accuracies on black defendants was 62.7%, and the average of the median accuracies on white defendants was 65.0%. The difference between these averages can be compared through a t-test for independent means. The samples are both independent, the data are approximately normally distributed, and the two samples have approximately the same variance, so the data passes the requirements to run this t-test. The test statistic was not significant at the 0.05 critical alpha level, t(19)=-1.11481, p=0.2719. Therefore, there is not sufficient evidence to suggest that the difference between the participants' accuracies on white versus black defendants is significant.

The crowd responses can also be analyzed for predictive parity. The crowd's accuracy on black defendants was 65.9%, and the crowd's accuracy on white defendants was 67.4%. A t-test for independent means was run on the 20 blocks of accuracies. The test statistic was not significant at the 0.05 critical alpha

level, t(19)=-0.46274, p=0.6462. Therefore, there is not sufficient evidence to suggest that the difference between the crowd's accuracies on black versus white defendants is significant.

The predictive parity of the participants and of the crowd can be compared to that of COMPAS on our subset of 1,000 defendants. The COMPAS accuracy for black defendants was 64.9%, and the COMPAS accuracy for white defendants was 65.7%. A t-test for independent means was run on the 20 blocks of accuracies for the defendants in the study. The test statistic was not significant at the 0.05 critical alpha level, t(19)= -0.25443, p=0.8005. Therefore, there is not sufficient evidence to suggest that the difference between the COMPAS accuracies on black versus white defendants is significant.

These results suggest that the individual participant predictions, the crowd predictions, and the COMPAS predictions satisfy predictive parity for white and black defendants, which means they are equally accurate for both groups.

Equalized Odds A prediction algorithm satisfies *equalized odds* if it has equal true positive rates and equal false positive rates for each group of individuals. This is equivalent to having equal false negative rates and equal false positive rates for each group.

Across the 400 participants, the average false positive rate for black defendants was 47.1% (median = 46.2%), and the average false positive rate for white defendants was 33.1% (median = 30.8%). These values can be compared through a t-test for independent means on the 20 independent block values. The average of the median false positive rates for black defendants

was 46.7%, and the average of the median false positive rates for white defendants was 30.9%. The test statistic was significant at the 0.05 critical alpha level, t(19)=5.8113, $p < 0.00001$. This suggests that the false positive rate for black defendants was significantly higher than for white defendants.

Across the 400 participants, the average false negative rate for black defendants was 31.5% (median = 29.4%), and the average false negative rate for white defendants was 41.4% (median = 42.3%). The average of the median false negative rates for black defendants was 29.8%, and the average of the median false negative rates for white defendants was 40.3%. The test statistic was significant at the 0.05 critical alpha level, t(19)=-2.4625, p=0.0184. This suggests that the false negative rate for black defendants was significantly lower than for white defendants. Therefore, these tests suggest that the individual participants' predictions do not satisfy equalized odds.

The crowd predictions can be similarly evaluated. In the crowd's predictions, the false positive rate for black defendants was 39.9%, and the false positive rate for white defendants was 26.2%. The test statistic for a t-test for independent means was significant at the 0.05 critical alpha level, t(19)=3.5236, p=0.0011. Therefore, there is sufficient evidence to suggest that the false positive rate for black defendants was significantly higher than for white defendants. The crowd's false negative rate for black defendants was 29.8%, and the false negative rate for white defendants was 43.6%. The test statistic was significant at the 0.05 critical alpha level, t(19)=-2.3295, p=0.0252. Therefore, there is sufficient evidence to suggest that the false negative rate for black defendants was significantly lower than for white defendants. Con-

sequently, these tests suggest that the crowd's predictions also do not satisfy equalized odds.

We can compare the participants' performance to the COMPAS algorithm. From the COMPAS predictions, the false positive rate for black defendants was 40.4%, and the false positive rate for white defendants was 25.4%. The test statistic for a t-test for independent means was significant at the 0.05 critical alpha level, t(19)=3.2452, p=0.0025. Therefore, there is sufficient evidence to suggest that the false positive rate for black defendants was significantly higher than for white defendants. The COMPAS false negative rate for black defendants was 30.9%, and the false negative rate for white defendants was 47.9%. The test statistic was significant at the 0.05 critical alpha level, t(19)=-3.1425, p=0.0032. Therefore, there is sufficient evidence to suggest that the false negative rate for black defendants was significantly lower than for white defendants. Thus, these tests suggest the COMPAS predictions also do not satisfy equalized odds.

These results suggest that the individual participant predictions, the crowd predictions, and the COMPAS predictions do not satisfy equalized odds, because they do not have equal true positive rates and equal false positive rates between white and black individuals. These results indicate that the human predictions are biased in the same ways as the COMPAS algorithm under this criterion of fairness.

Accuracy Equity Recidivism risk scores satisfy *accuracy equity* if they can discriminate recidivists and non-recidivists equally well for two different groups

[15]. A risk score's discriminative ability is measured by its AUC-ROC value. For the crowd predictions, the percentage of the crowd that answered "Yes" on a defendant represents the defendant's "risk score". The AUC-ROC value for the crowd's predictions on black defendants was 0.6902, and the AUC-ROC value for the crowd's predictions on white defendants was 0.7127.

The significance of the difference between the AUC-ROC values was computed [19]. The test statistic was not significant at the 0.05 critical alpha level, z(1)=-0.6203, p=0.5351. Therefore, there is not sufficient evidence to suggest that the AUC-ROC values for the black and white defendants are significantly different. This result suggests that the crowd risk scores satisfy accuracy equity for black and white defendants.

For the COMPAS scores, the AUC-ROC value for black defendants was 0.6906, and the AUC-ROC value for white defendants was 0.6937. The test statistic was not significant at the 0.05 critical alpha level, z(1)=-0.0845, p=0.9327. Therefore, there is not sufficient evidence to suggest that the AUC-ROC values for the black and white defendants are significantly different. This result suggests that the COMPAS risk scores also satisfy accuracy equity for black and white defendants.

These results indicate that the crowd predictions are equivalently discriminative to the COMPAS algorithm.

Calibration A risk score is *well-calibrated* by race if each score has the same meaning regardless of a defendant's race. Therefore, white and black defendants with the same risk score should be equally likely to re-offend. The

crowd risk scores, determined by the percentage of the crowd that answered "Yes" on a defendant, can be binned into 10 distinct risk scores from 1-10. Shown in Figure 2.6 and Figure 2.7 are the percentage of individuals that recidivated at each of the crowd risk scores and COMPAS risk scores for both black and white defendants.

At each crowd score, the 95% confidence intervals for the percentage of black individuals who recidivated and the percentage of white individuals who recidivated are overlapping. This suggests that white and black defendants with the same crowd risk score are equally likely to reoffend.

The COMPAS scores, however, are slightly less calibrated. The 95% confidence intervals do not overlap at all of the scores, suggesting that white and black defendants with the same COMPAS risk score are not always as equally likely to reoffend. It is important to note, however, that the COMPAS risk scores are well-calibrated on the entire data set of defendants.

These results suggest that the crowd predictions are at least as well-calibrated, and possibly more calibrated, than the COMPAS predictions.

2.3 Summary

Our study participants were able to achieve a median accuracy of 64.0%, and their aggregated crowd responses produced a prediction accuracy of 66.5%. These prediction accuracies are not significantly different from the COMPAS accuracy of 65.2%. The crowd predictions resulted in an AUC-ROC value of

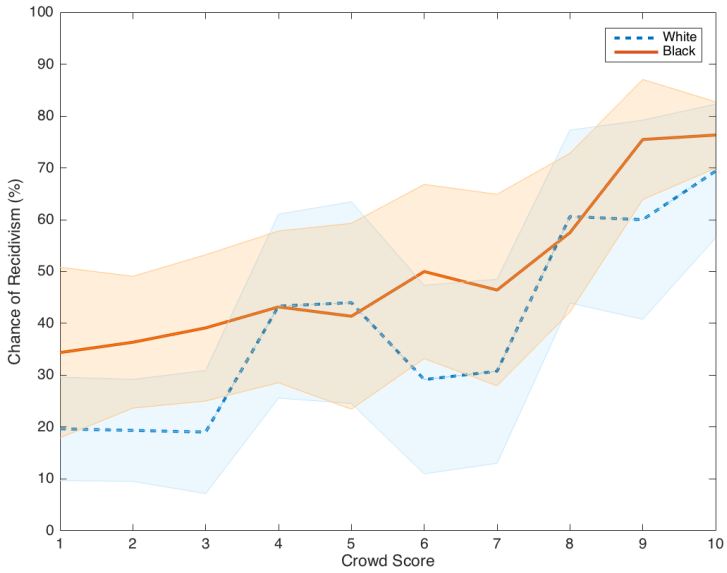

Figure 2.6: Recidivism rate by crowd risk score. The shaded areas represent 95% confidence intervals at each value.

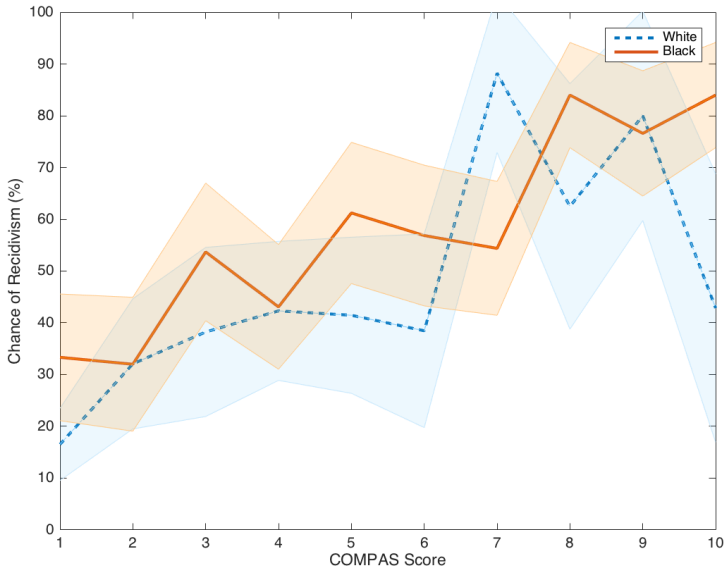

Figure 2.7: Recidivism rate by COMPAS risk score. The shaded areas represent 95% confidence intervals at each value.

0.71, which is statistically comparable to the COMPAS AUC-ROC value of 0.70.

When evaluated in terms of four main fairness criteria (predictive parity, equalized odds, accuracy equity, and calibration) both human prediction and COMPAS prediction fully satisfied predictive parity and accuracy equity, and both were well-calibrated. Neither human prediction nor COMPAS prediction satisfied the equalized odds fairness criterion that expects equal false positive and false negative rates across groups. Therefore, the participants performed with both equivalent accuracy and equivalent fairness to the COMPAS algorithm.

2.4 Human Recidivism Prediction Without Race

COMPAS scores are calculated from 137 characteristics of a defendant. Race is not included in these characteristics [5]. The Mechanical Turk participants, however, were shown the race of the defendants to inform their predictions. Consequently, there is a possibility that seeing the race of a defendant affected recidivism predictions. A new study was run to determine if the race of a defendant affected the participant predictions of their recidivism. The goal of this new study is to determine if the participant accuracy could be maintained while minimizing the equalized odds fairness violation.

2.4.1 Methods

This study used the same 1,000 defendants from the first study. The description paragraphs were identical, except the race of the defendant was removed. Thus, the participants were shown the following details of each defendant:

1. Sex
2. Age
3. Criminal Charge Description
4. Criminal Charge Degree
5. Juvenile Misdemeanor Count
6. Juvenile Felony Count
7. Number of Prior Crimes

Shown in Figure 2.8 is an example question from the new study, with the race of the defendant removed (see Figure 2.1). Besides removing the race from the questions, this study was conducted in the same way as the first study. However, this study had different participants from the first study.

2.4.2 Results

The 400 participants correctly predicted criminal recidivism with a median accuracy of 64.0%, and a mean accuracy of 62.1%. The distribution of the 400 participant accuracies is shown in Figure 2.9.

The accuracy of the participants was analyzed in comparison to the first study that included the race of the defendant. Across the 20 blocks, the average

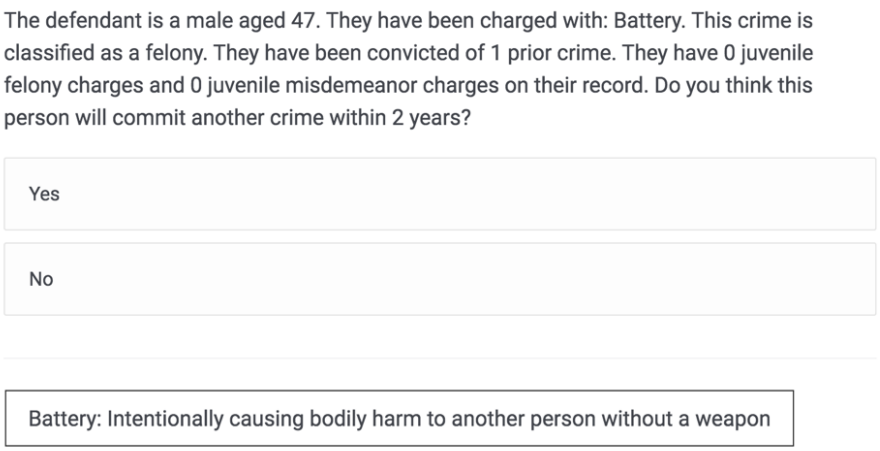

Figure 2.8: Sample question with the race removed. See also Figure 2.1.

of the 20 median participant accuracies was 63.1% for the study with race, and 62.8% for the study without race. A matched-pairs t-test was performed to determine if the difference was significant. The test statistic was not significant at the 0.05 critical alpha level, t(19)=0.5325, p=0.6005. Therefore, there is not sufficient evidence to suggest that the accuracy from the study with race is significantly different from the study without race.

The crowd performance was analyzed in comparison to the COMPAS performance on the 1,000 defendants. Across the 20 blocks, the average crowd accuracy for the study with race was 66.5%, and the average crowd accuracy for the study without race was 67.0%. A matched-pairs t-test was performed to determine if the difference between those accuracies was significant. The test statistic was not significant at the 0.05 critical alpha level, t(19)=-0.4453, p=0.6611.

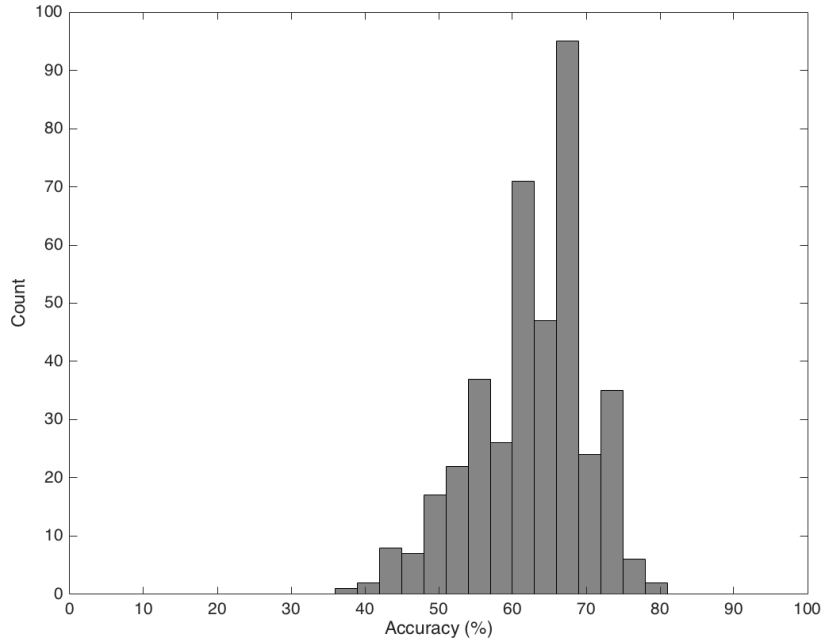

Figure 2.9: Participant accuracy distribution for study without race.

The crowd performance was also assessed according to its AUC-ROC value. The AUC-ROC value of the crowd was 0.7100, indicating satisfactory predictive accuracy. The AUC-ROC value for the crowd with race was 0.7101. Shown in Figure 2.10 are the overlapping ROC curves of the crowds with and without race.

The individual and crowd performances on the black and white defendants from each study can be directly compared using matched-pairs t-tests of the 20 blocks to determine if the absence of race affected the participants' ability to predict recidivism for either race. The accuracy, false positive, and false negative rates for both black and white defendants were compared between the

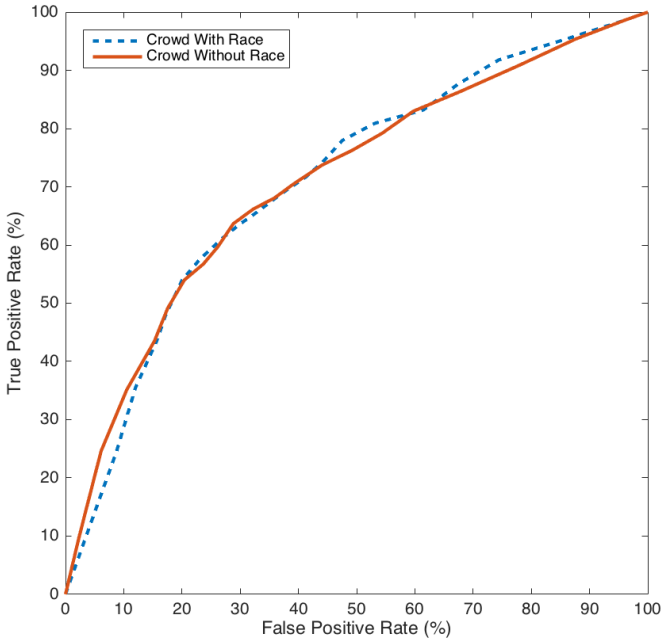

Figure 2.10: ROC curves of the crowd from the study with race and the crowd from the study without race.

two studies. No values from one study were statistically significantly different from that of the other study. Therefore, there is not sufficient evidence to suggest that the participant and crowd performance by race was significantly different between the two studies. The details of this analysis can be found in Appendix E.

Fairness

We can also evaluate the fairness of the participants in the study without race in terms of how well their results meet the same fairness criteria used to

evaluate the study with race: Predictive Parity; Equalized Odds; Accuracy Equity; and Calibration.

Predictive Parity An algorithm satisfies *predictive parity* if it is equally accurate across all groups [4]. Therefore, the participants satisfy predictive parity if their responses were equally accurate for white and black defendants. Across the 400 individual participant responses, the average accuracy on black defendants was 62.7% (median = 62.5%), and the average accuracy on white defendants was 62.5% (median = 62.5%). Because the individual participant accuracies aren't independent, they must be analyzed by independent block accuracies. Therefore, one can compare the accuracies of the participants on black defendants in each block to the accuracies on white defendants in each block. Again, the median accuracies for each block were used. The average of the median accuracies on black defendants was 63.8%, and the average of the median accuracies on white defendants was also 63.8%. The difference between these averages can be compared through a t-test for independent means. The test statistic was not significant at the 0.05 critical alpha level, t(19)=-0.0061, p=0.9952. Therefore, there is not sufficient evidence to suggest that the difference between the participants' accuracies on white versus black defendants is significant.

The crowd responses can also be analyzed for predictive parity. The crowd's accuracy on black defendants was 68.2%, and the crowd's accuracy on white defendants was 67.6%. A t-test for independent means was run on the 20 blocks of accuracies. The test statistic was not significant at the 0.05 critical

alpha level, t(19)=0.1701, p=0.8658. Therefore, there is not sufficient evidence to suggest that the difference between the crowd's accuracies on black versus white defendants is significant.

The results of these two tests suggest that both the individual participant predictions and the crowd predictions satisfy predictive parity for white and black defendants, which means they are equally accurate for both groups.

Equalized Odds A prediction algorithm satisfies *equalized odds* if it has equal true positive rates and equal false positive rates for each group of individuals. This is equivalent to having equal false negative rates and equal false positive rates for each group. Across the 400 participants, the average false positive rate for black defendants was 44.9% (median = 44.4%), and the average false positive rate for white defendants was 35.3% (median = 35.7%). These values can be compared through a t-test for independent means on the 20 independent block values. The average of the median false positive rates for black defendants was 43.9%, and the average of the median false positive rates for white defendants was 33.8%. The test statistic was significant at the 0.05 critical alpha level, t(19)=3.0025, p=0.0047. This suggests that the false positive rate for black defendants was significantly higher than for white defendants.

Across the 400 participants, the average false negative rate for black defendants was 32.5% (median = 30.8%), and the average false negative rate for white defendants was 40.9% (median = 40.0%). The average of the median false negative rates for black defendants was 31.5%, and the average of the

median false negative rates for white defendants was 39.9%. The test statistic was not significant at the 0.05 critical alpha level, t(19)=-2.0170, p=0.0508. Therefore, there is not sufficient evidence to suggest that the false negative rate for black defendants was significantly lower than for white defendants. These tests suggest that while the individual participants' predictions do not satisfy the equal false positive rate criterion for equalized odds, they do satisfy the equal false negative rate criterion.

The crowd predictions can also be evaluated through this test. In the crowd's predictions, the false positive rate for black defendants was 37.1%, and the false positive rate for white defendants was 27.2%. The test statistic for a t-test for independent means was significant at the 0.05 critical alpha level, t(19)=2.2959, p=0.0273. Therefore, there is sufficient evidence to suggest that the false positive rate for black defendants was significantly higher than for white defendants. The crowd's false negative rate for black defendants was 29.2%, and the false negative rate for white defendants was 40.2%. The test statistic was significant at the 0.05 critical alpha level, t(19)=-2.1942, p=0.0344. Therefore, there is sufficient evidence to suggest that the false negative rate for black defendants was significantly lower than for white defendants. Consequently, these tests suggest that the crowd's predictions do not satisfy equalized odds.

These results suggest that the individual participant predictions and the crowd predictions do not fully satisfy equalized odds. However, the individual participant predictions from this study are closer to satisfying equalized odds than from the previous study, because these individual participant predictions

did not produce statistically different false negative rates for white and black defendants.

Accuracy Equity Recidivism risk scores satisfy *accuracy equity* if they can discriminate recidivists and non-recidivists equally well for two different groups [15]. A risk score's discriminative ability is measured by its AUC-ROC value (the area under the receiver operating characteristic (ROC) curve). The AUC-ROC value for the crowd's predictions on black defendants was 0.6993, and the AUC-ROC value for the crowd's predictions on white defendants was 0.7035. Therefore, the crowd was satisfactorily discriminative for both black and white defendants.

The significance of the difference between the AUC-ROC values was computed [19]. The test statistic was not significant at the 0.05 critical alpha level, z(1)=-0.1156, p=0.9080. Therefore, there is not sufficient evidence to suggest that the AUC-ROC values for the black and white defendants are significantly different. This result suggests that the crowd risk scores satisfy accuracy equity for black and white defendants.

Calibration A risk score is *well-calibrated* by race if each score has the same meaning regardless of a defendant's race. Therefore, white and black defendants with the same risk score should be equally likely to reoffend. Shown in Figure 2.11 is the percentage of individuals that recidivated at each of the crowd risk scores for both black and white defendants.

At all scores except for 1 and 9, the 95% confidence intervals for the percentage of black individuals who recidivated and the percentage of white indi-

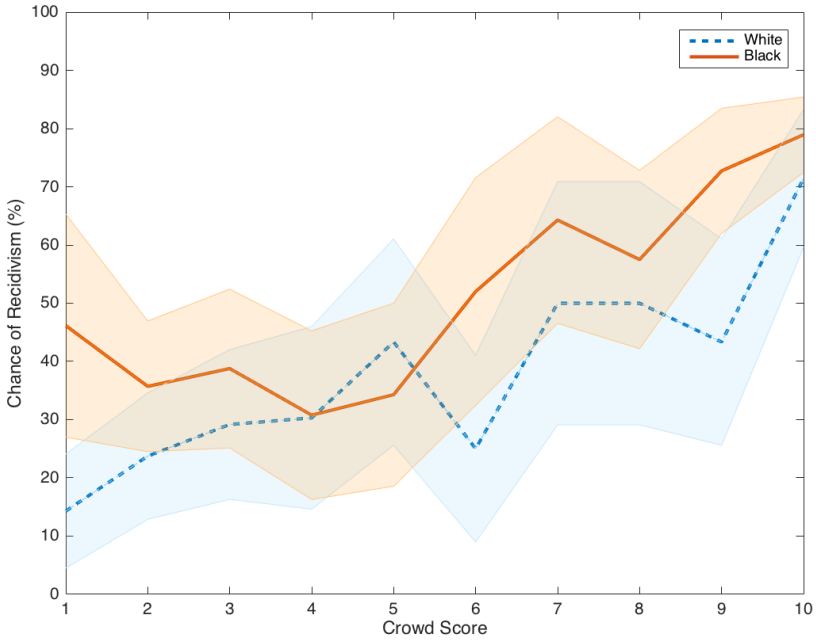

Figure 2.11: Recidivism rate by crowd risk score. The shaded areas represent 95% confidence intervals at each value.

viduals who recidivated are overlapping. This suggests that white and black defendants with the same crowd risk score are equally likely to reoffend for every score except for 1 and 9. This indicates that the crowd predictions are almost entirely well-calibrated, but are slightly less calibrated than in the study that included the defendant's race.

2.4.3 Summary

When not informed of a defendant's race, the study participants were able to achieve a median accuracy of 64.0%, and their aggregated crowd responses produced a prediction accuracy of 67.0%. The participants of the study that

included the defendant's race also achieved a median accuracy of 64.0%, and their aggregated crowd responses produced a prediction accuracy of 66.5%. The prediction accuracies from the study without race are not significantly different from those of the study that included the defendants race. The crowd predictions in this study resulted in an AUC-ROC value of 0.71, which is equivalent to the crowd's AUC-ROC value of 0.71 for the study that included race. The participants' performance by race was also equivalent to that of the previous study.

When evaluated in terms of four main fairness criteria (predictive parity, equalized odds, accuracy equity, and calibration), both the individual participant and the crowd predictions satisfied predictive parity and accuracy equity in the same ways as in the first study. However, the individual participant predictions from the study without race were closer to satisfying equalized odds than the predictions from the study with race. Additionally, the crowd scores from this study were slightly less calibrated than the first study, but still satisfied the calibration criterion for eight of the ten risk scores. This is the same level of calibration that COMPAS achieves on this subset of defendants.

2.5 Discussion

The results of the initial study indicate that human predictions perform comparably to the COMPAS algorithm. The participants achieved equivalent accuracy to COMPAS via both individual predictions and aggregated crowd predictions. The crowd predictions could discriminate between recidivists and

non-recidivists with the same consistency as the COMPAS algorithm, as well. These results suggest that there is no predictive advantage to using the COMPAS algorithm over simple human prediction when estimating a defendant's risk of recidivism.

Algorithmic presence in the criminal justice system is often justified due to its presumed objectivity in comparison to human prediction. This study found that COMPAS and human predictions satisfied the same fairness qualifications, and suffered from the same type of bias. Both predictions satisfied predictive parity, accuracy equity, and were well-calibrated. Both failed to satisfy equalized odds in the same way; the false positive rate for black individuals was significantly higher than for white individuals, and the false negative rate for white individuals was significantly higher than for black individuals. However, it is essential to recognize that satisfying all four fairness criteria is not mathematically possible given the difference in base recidivism rates between the white and black defendants. Therefore, while both COMPAS and the human predictions fail to satisfy equalized odds, neither could possibly satisfy this criterion without violating the calibration criterion. Overall, this study suggests that the COMPAS algorithm does not offer any form of objectivity that is lacking in humans.

One limitation of our study is the racial homogeneity of the participants. Of the 400 Mechanical Turk participants in the initial study, 77.3% of them identified as white. Only 7.5% of the participants identified as black. Consequently, the race of the participants may have affected how they predicted recidivism. Because there were so few non-white participants, the racial bi-

ases of participants were not analyzed by their identified race. However, when considering these results as indicative of human recidivism prediction ability and of human racial biases, it is important to acknowledge that these results are specific to a mostly white population. Opportunities for future research could include diversifying the racial demographics of the human participants, and analyzing how one's race affects their biases in recidivism prediction.

Overall, these results indicate that humans can predict recidivism in a way that is equivalent to the COMPAS algorithm in terms of both accuracy and fairness. It is important to note that the COMPAS algorithm uses 137 features to predict recidivism, while the human participants were given only 8 features for each defendant. The participants successfully matched the accuracy of COMPAS with much less information about each defendant. Future research on this topic could explore whether humans can be more predictive than COMPAS if given more details on each defendant.

The results of the follow-up study provide insight into the role that the race of a defendant plays in human recidivism prediction. Although the participants were not shown the race of a defendant in this second study, they performed with comparable accuracy to the study that included the defendant's race. The crowd predictions were also comparably accurate to the crowd predictions of the study that included the defendant's race. These results suggest that knowing the race of a defendant does not affect a person's accuracy or bias in predicting the defendant's risk of recidivism.

While the overall performance of the participants in the second study was equivalent to that of the first study, the fairness performance of the predictions

in the second study was slightly different than in the first study. Hiding the defendant's race produced false negative rates for white and black individuals that were not significantly different from each other. This result suggests humans are more likely to incorrectly predict a white defendant will not recidivate when the participant knows the defendant is white. Conversely, this result suggests that humans are less likely to incorrectly predict a black defendant will not recidivate when the participant knows the defendant is black. This finding implies that white defendants benefit from their race being known, whereas black defendants suffer from having their race known.

This study also set out to determine a baseline accuracy of human recidivism prediction. The median accuracy of both studies was 64.0%. However, the participants involved in the study were untrained individuals that have presumably no background in criminal justice or criminal behavior. Consequently, it is reasonable to expect that people with more experience in this field could perform with an even higher accuracy than these untrained participants. Thus, while an accuracy of 64.0% can be understood as an estimate of the baseline human accuracy, it should be noted that this may be a conservatively low estimate of the potential for trained human recidivism prediction. A worthy corollary for future research would be replicating this study with a set of participants who have experience studying recidivism.

3. Algorithmic Recidivism Prediction

One major qualm about algorithmic predictions is that they are considered black boxes [12, 14, 32, 29], which is a term used to describe a machine that takes a series of inputs and produces results from a secret, unexplained process [29]. Companies that offer algorithmic prediction services profit from the tools they develop, so the inner workings of their algorithms are considered intellectual property. A particular worry about using recidivism prediction instruments is that judges who see these risk scores aren't aware of how the various components of a defendant's identity (demographics, criminal history, etc.) are being used to predict recidivism [32]. While Northpointe does not share how, exactly, their COMPAS algorithm works, we can build our own classifiers in an attempt to replicate the behavior and performance of COMPAS. Furthermore, replicating the COMPAS algorithm provides an opportunity to explore whether incorporating human judgment into algorithms can enhance prediction accuracy.

3.1 Replicating the COMPAS Algorithm

We built multiple classifiers with the goal of reproducing, and possibly exceeding, the prediction accuracy of COMPAS. The classifiers were trained using the following eight features, provided by the data set published by Angwin et al. in the ProPublica study [25].

1. Sex
2. Age
3. Race
4. Criminal Charge Degree (Misdemeanor or Felony)
5. Juvenile Misdemeanor Count
6. Juvenile Felony Count
7. Juvenile "Other" Charges Count
8. Number of Prior Crimes

Using these features, Linear Discriminant Analysis and Logistic Regression classifiers were explored.

Linear Discriminant Analysis (LDA) Linear Discriminant Analysis (LDA) is a linear classifier that is used to discriminate between two or more classes of data. LDA works by projecting the data onto a linear subspace so that the within-class variances are minimized, and the between-class variance is maximized. In the binary prediction case, this method assumes that the

data come from two Gaussian distributions, each with different means but with the same covariance [16].

The LDA classifier was trained on a random subset of 80% of the data, and then tested on the remaining 20%. This process was repeated 1,000 times, with a new random training and testing split each time. The average test accuracy was 63.0% (σ=2.44). The average training accuracy was 63.1% (σ=0.67). The results exhibit qualitatively the same disparities between the false positive and false negative rates for white and black defendants as COMPAS, Table 3.1.

Logistic Regression Logistic Regression learns a sigmoidal mapping that maximizes the log-likelihood of each training example being correctly classified. Logistic regression produces probability estimates of a test example belonging to each class. Unlike LDA, this method does not assume a particular distribution of the data [28].

The Logistic Regression classifier was also trained on a random subset of 80% of the data, and then tested on the remaining 20%. This process was repeated 1,000 times, with a different training and testing split each time. The average test accuracy was 66.9% (σ=1.16). The average training accuracy was 67.0% (σ=0.35). The results exhibit qualitatively the same disparities between the false positive and false negative rates for white and black defendants as COMPAS, Table 3.1.

Summary The performance of each classifier was compared to the performance of COMPAS on the full set of defendants in the data set. Shown in

Method	Defendant Category	Accuracy (%)	False Positive Rate (%)	False Negative Rate (%)
COMPAS	**Overall**	65.3	32.4	37.5
	Black	63.8	44.9	28.1
	White	66.9	23.5	47.9
LDA	**Overall**	63.0	36.9	37.1
	Black	63.3	45.8	28.5
	White	62.3	29.1	51.3
LR	**Overall**	66.9	32.8	33.5
	Black	67.2	42.9	23.3
	White	65.7	26.8	45.8

Table 3.1: Performance of COMPAS, Linear Discriminant Analysis, and Logistic Regression.

Table 3.1 are the accuracy, false positive rate, and false negative rate values for each classification method, including COMPAS.

While none of these classifiers significantly exceeded the predictive ability of COMPAS, we were able to successfully reproduce the accuracy of COMPAS. Both of these linear classifiers performed similarly to COMPAS on this data set, which suggests that COMPAS employs a linear classifier in their product. These classifiers were trained on only eight variables for each defendant, whereas COMPAS classifiers are trained on 137 [4]. These equivalent classifiers suggest that the variables that Northpointe uses are no more predictive than these eight features.

3.2 Improving the COMPAS Algorithm

Beyond replicating the COMPAS accuracy, we explored whether incorporating human judgment into the classifier could increase the accuracy. This analysis incorporated the human crowd predictions on each of 1,000 defendants. The

crowd predictions from the original study with race were used. As previously discussed, these crowd predictions can be translated into risk scores by calculating the percent of participants that answered "Yes" for a defendant. These 1,000 defendants also have risk scores assigned by COMPAS. We combined the crowd risk scores, the COMPAS risk scores, and the demographic and criminal history information of these 1,000 defendants to determine if we could build a more accurate classifier using both human and algorithmic knowledge. The following features were used to build the classifiers:

1. Crowd Risk Score
2. COMPAS Recidivism Risk Score
3. Sex
4. Age
5. Race
6. Criminal Charge Degree (Misdemeanor or Felony)
7. Juvenile Misdemeanor Count
8. Juvenile Felony Count
9. Juvenile "Other" Charges Count
10. Number of Prior Crimes

These features were used to re-build the LDA and LR classifiers. They were also used to build a non-linear Support Vector Machine. We included the COMPAS scores to simulate incorporating COMPAS scores and human judgment. The additional demographic and criminal history variables were

also used to further boost the prediction accuracy. Because the crowd risk scores were utilized, these classifiers were trained and tested on only the 1,000 defendants that we have crowd risk scores for from the study.

Linear Discriminant Analysis The LDA classifier was re-built using this new set of features. It was trained on a random subset of 80% of the data, and then tested on the remaining 20%. This process was repeated 1,000 times, with a new random training and testing split each time. The average test accuracy was 61.7% (σ=3.44). The average training accuracy was 61.5% (σ=1.54). Including the crowd and COMPAS scores did not significantly change the results of this classifier, Table 3.2.

Logistic Regression The LR classifier was re-built using this new set of features. It was trained on a random subset of 80% of the data, and then tested on the remaining 20%. This process was repeated 1,000 times, with a new random training and testing split each time. The average test accuracy was 66.3% (σ=2.92). The average training accuracy was 67.1% (σ=0.90). Including the crowd and COMPAS scores did not significantly change the results of this classifier, Table 3.2.

Non-Linear Support Vector Machine A Support Vector Machine (SVM) is a classifier that fits an optimal separating hyperplane to the labeled training data. The optimal separating hyperplane is found by maximizing the margin between the hyperplane and the closest points to the plane in each class. Non-Linear Support Vector Machines handle classification problems that do

not have a linear separating hyperplane. These SVMs handle non-linearly separable cases by raising the dimensionality of the features through a kernel function that separates the data in a higher dimension [8]. A radial basis function kernels was implemented for our classifier.

The Non-Linear SVM was trained on a random subset of 80% of the data, and then tested on the remaining 20%. This process was repeated 1,000 times, with a new random training and testing split each time. The average test accuracy was 67.1% (σ=2.98), which is almost exactly the same as the average accuracy for the linear SVM. The average training accuracy was 75.3% (σ=0.82). These results also suffer from the same false positive and false negative rate disparities between white and black defendants, Table 3.2.

Summary These results were compared to the COMPAS performance on the subset of 1,000 defendants in the study. Shown in Table 3.2 are the accuracy, false positive rate, and false negative rate values for COMPAS and these three classifiers.

While these classifiers were also able to replicate the performance of COMPAS, they did not perform significantly better than COMPAS. This suggests that incorporating human judgment into the classifiers does not significantly increase the accuracy.

3.3 Discussion

We successfully built two linear classifiers that replicated the behavior and performance of COMPAS. The ability of LDA and LR to reproduce the re-

Method	Defendant Category	Accuracy (%)	False Positive Rate (%)	False Negative Rate (%)
COMPAS	**Overall**	65.2	31.5	38.5
	Black	64.9	41.2	30.5
	White	65.8	24.9	50.0
LDA	**Overall**	61.7	38.1	38.4
	Black	61.3	50.0	33.1
	White	62.1	32.7	46.7
LR	**Overall**	66.3	32.1	35.1
	Black	66.1	46.9	24.1
	White	67.3	24.5	46.5
Non-Linear SVM	**Overall**	67.1	22.3	44.4
	Black	66.2	30.6	36.0
	White	68.8	15.5	57.4

Table 3.2: Performance of COMPAS, LDA, LR, and Non-Linear SVM.

sults of COMPAS suggests that Northpointe uses a linear classifier to make their predictions. At the very least, even if Northpointe employs a more complicated algorithm than a linear classifier, their performance does not exceed that of a simple linear classifier. Furthermore, the eight variables used in our classifiers were equally as predictive as Northpointe's 137 variables. This result implies that resources are wasted collecting the data for the 137 questions that Northpointe demands, for the eight variables that ProPublica was able to easily obtain are equally predictive.

Adding human judgment did not significantly affect the performance of the LDA and LR classifiers. While COMPAS and the crowd were equally predictive on their own, combining them did not increase their overall prediction accuracy. This suggests that crowd wisdom about recidivism does not enhance algorithmic accuracy.

The non-linear SVM did not provide any predictive advantage over the linear classifiers. This lack of classification improvement suggests that the data

may not be separable. In other words, the inability of the non-linear SVM to predict with high accuracy suggests that these features are not discriminatory in predicting future recidivism. Furthermore, these features are equivalently predictive to the 137 features that COMPAS uses. Thus, more information about a defendant also does not necessarily guarantee a higher prediction accuracy.

4. Conclusion

The use of COMPAS in the criminal justice system has been challenged in the past year because it is a secret "black box", and its predictions cannot be contested in a courtroom in the same way other claims are refuted [26]. Trusting a "black box" system in the courtroom can possibly be justified if its predictions are substantially more accurate than human judgment. However, this study shows that COMPAS is not significantly more accurate than untrained human predictions. Additionally, COMPAS could be a worthy tool in the courtroom if it offered a form of objectivity of which humans aren't capable. However, in terms of racial biases, this study exposed that COMPAS does not offer any fairness advantages over human prediction.

The second study revealed that knowing the race of a defendant does not significantly affect a person's accuracy in predicting the defendant's risk of recidivism. This finding could have serious implications in the criminal justice system. Qualms about human judgment in criminal justice are often based in fears of people's intentional and unintentional racial biases. If knowing a defendant's race does not significantly improve a person's ability to predict that defendant's risk of recidivism, then conducting race-blind processes may help mitigate racial prejudice in the criminal justice system.

Replicating the COMPAS algorithm contributed an essential perspective of recidivism prediction. Knowing more information about a defendant does not

necessarily result in more accurate predictions. Additionally, building more complex, non-linear classifiers does not enhance predictive ability, suggesting that demographic and criminal history information is not predictive of recidivism.

The findings of this study could have significant consequences for the realm of recidivism prediction. The continued use of COMPAS in courtrooms should be called into question, for it is neither more accurate nor less biased than human judgment. Because the base rates of recidivism are different between white an black defendants, some type of bias is mathematically inevitable in recidivism prediction, regardless of the prediction mechanism. Furthermore, our findings indicate that neither more information nor more complex algorithms can enhance recidivism predictions. Therefore, if recidivism predictions are inevitably biased and only moderately predictive, then perhaps recidivism is not something worth trying to predict.

Acknowledgments

I would like to thank Professor Hany Farid for his wisdom and guidance throughout this project. He has dedicated countless hours to advising me on this thesis, and I am incredibly appreciative of his enthusiastic support over the past year. I would also like to thank Lauren Lax for her critical advice on the statistical components of this project.

A. Crime Descriptions

Crime descriptions as described in Florida laws [23]. These crime descriptions appearred in the questions of the study to clarify the particular crime with which a defendant was charged.

1. Abuse: The intentional infliction of physical or psychological injury
2. Armed Robbery: When a person intentionally and unlawfully takes money or property from another person through the use of force, violence, assault, or threat while in possession of a firearm or other deadly weapon
3. Assault: An intentional and unlawful threat against another person, coupled with the apparent ability to carry out the threat, which creates a genuine and reasonable fear that violence or harm is imminent
4. Assault with a Deadly Weapon: An intentional and unlawful threat against another person with a deadly weapon, or while in the commission of a felony, which creates a reasonable fear that violence or harm is imminent
5. Battery: Intentionally causing bodily harm to another person without a weapon

6. Battery with a Deadly Weapon: Intentionally causing bodily harm to another person with a deadly weapon

7. Burglary: Unlawfully entering a dwelling or structure or remaining in a dwelling or structure after permission to remain has been withdrawn, with the intent to commit a crime inside

8. Carrying a Concealed Weapon: When a person knowingly carries, on or about their person, a weapon that is concealed from the ordinary sight of another person

9. Child Abuse: Intentional infliction of physical or mental injury upon a child

10. Child Molestation: Intentionally touching the breasts, genitals, or buttocks of a child younger than 16 in a lewd or lascivious manner

11. Child Neglect: When a caregiver willfully, or through culpable negligence, fails to provide a child with the care, supervision, and services necessary to maintain the child's physical and mental health that a prudent person would consider essential for the well-being of the child

12. Contributing to the Delinquency Of A Minor: When a person over the age of eighteen commits any act which causes, tends to cause, encourages, or contributes to a minor become delinquent, dependent, or a child in need of services

13. Criminal Damage of less than $1,000: When a person willfully and maliciously damages another person's property - with damage less than $1,000

14. Criminal Damage of more than $1,000: When a person willfully and maliciously damages another person's property - with damage of more than $1,000

15. Dealing Cannabis/Marijuana: Possessing cannabis with the intent to sell or deliver the cannabis

16. Dealing Cocaine: Possessing cocaine with the intent to sell or deliver the cocaine

17. Dealing Controlled Substances: Possessing a criminalized narcotic or pharmacological drug with the intent to sell or deliver

18. Disorderly Conduct: Committing an act that corrupts the public morals, outrages the sense of public decency, or affects the peace and quiet of persons who may witness them, or engages in brawling, fighting, or other conduct that constitutes a breach of the peace

19. Disorderly Intoxication: When an intoxicated person endangers the safety of another person or property, or causes a disturbance in public

20. Domestic Violence: The touching or striking of a family member, household member, or domestic partner against their will

21. Driving Under the Influence: Operating a motor vehicle with a blood alcohol content (BAC) of 0.08% or higher

22. Driving with a Revoked License: Driving with a license that has been revoked

23. Driving with a Suspended License: Driving with a license that has been suspended

24. Driving with an Expired License: Driving with a license that is expired

25. Drug Trafficking: To sell, purchase, manufacture, deliver, possess, or transport a trafficking amount (i.e., more than personal use) of drugs

26. Escape: When a prisoner escapes or attempts to escape from a place of confinement, or when an arrested person who is being transported to or from a place of confinement escapes or attempts to escape from such lawful confinement

27. Extradition of Defendants: The transfer of an accused criminal by one state or nation to another

28. Failure to Obey Police Officer: Failure to obey the orders of a police officer

29. False Imprisonment: When a person either forcibly, by threat, confines, abducts, imprisons, or restrains another person without lawful authority against their will, or secretly confines, abducts, imprisons, or restrains another person without lawful authority against their will

30. Fleeing the Scene of an Accident: When a person is involved in an accident or crash and willfully leaves the scene of the accident or crash without providing their information to the other individuals involved

31. Forgery: When a person falsely makes, alters, counterfeits, or forges a document that carries legal efficacy

32. Fraud: Wrongful or criminal deception intended to result in financial or personal gain

33. Grand Theft: The unlawful taking of property worth more than $300

34. Interference with Traffic Control Railroad Divide: Unlawfully interfering with a railroad divide

35. Kidnapping: When a person forcibly, secretly, or by threat confines, abducts, or imprisons another person against their will, without lawful authority

36. Loitering: Standing or waiting around idly in a manner unusual for law-abiding citizens, creating an imminent threat to public safety

37. Manufacturing Cannabis/Marijuana: Manufacturing or cultivating cannabis

38. Offense Against Intellectual Property: Robbing people or companies of their ideas, inventions, and creative expressionsknown as intellectual propertywhich can include everything from trade secrets and proprietary products and parts to movies, music, and software

39. Open Carrying of Weapon: When a person knowingly carries, on or about his or her person, a weapon that is not concealed

40. Operating a Vehicle without a Valid Drivers License: Operating a motor vehicle without a valid drivers license

41. Possession of a Controlled Substance: Possession of a criminalized narcotic or pharmacological drug

42. Possession of Cannabis/Marijuana: Possession of cannabis/marijuana

43. Possession of Cocaine: Possession of cocaine

44. Possession of Ecstasy: Possession of Ecstasy

45. Possession of LSD: Possession of LSD

46. Possession of Meth: Possession of Meth

47. Possession of Morphine: Possession of Morphine

48. Possession of Oxycodone: Possession of Oxycodone

49. Prostitution: Engaging in sex for money

50. Reckless Driving: When you drive a vehicle in a manner that shows a willful disregard for the safety of persons or property

51. Resisting an Officer: When a person knowingly and willfully, but not violently, resists, obstructs, or opposes a law enforcement officer engaged in the execution of legal process, or lawful execution of a legal duty

52. Resisting an Officer with Violence: When a person knowingly and willfully resists, obstructs, or opposes a law enforcement officer by threatening violence or engaging in violent conduct against the law enforcement officer was engaged in the lawful execution of a legal duty

53. Restraining Order Violation: Violating the terms of a restraining order

54. Robbery: When a person intentionally and unlawfully takes money or property from another person through the use of force, violence, assault, or threat

55. Sexual Assault: Sexual assault, also known as rape, sexual abuse and sexual battery, is defined as unwanted oral, anal or vaginal penetration by, or union with, the sexual organ of another or the anal or vaginal penetration of another by any other object

56. Soliciting For Prostitution: When a person solicits, induces, entices, or procures another person to engage in prostitution, lewdness, or assignation

57. Stalking: When a person willfully, maliciously, and repeatedly follows, harasses, or cyberstalks another person

58. Tampering with a Witness: The act of attempting to alter or prevent the testimony of witnesses within criminal or civil proceedings

59. Tampering With Physical Evidence: When a person alters, destroys, suppresses or conceals any record, document or thing with purpose to

impair its verity, legibility or availability in any official proceeding or investigation

60. Theft: The taking of another person's property worth less than $300

61. Threat Against Public Servant: An intentional and unlawful threat against an emergency medical care provider, firefighter, or law enforcement officer, coupled with the apparent ability to carry out the threat, which creates a genuine and reasonable fear that violence or harm is imminent

62. Trespassing: Willfully entering or remaining on some form of real property without authorization, license, or invitation

63. Unlicensed Telemarketing: When an unlicensed commercial telemarketing company or unlicensed salesperson solicits a purchaser for the purpose of attempting to sell consumer goods or services

B. Study Demographic Questions

Each participant answered the following demographic questions.

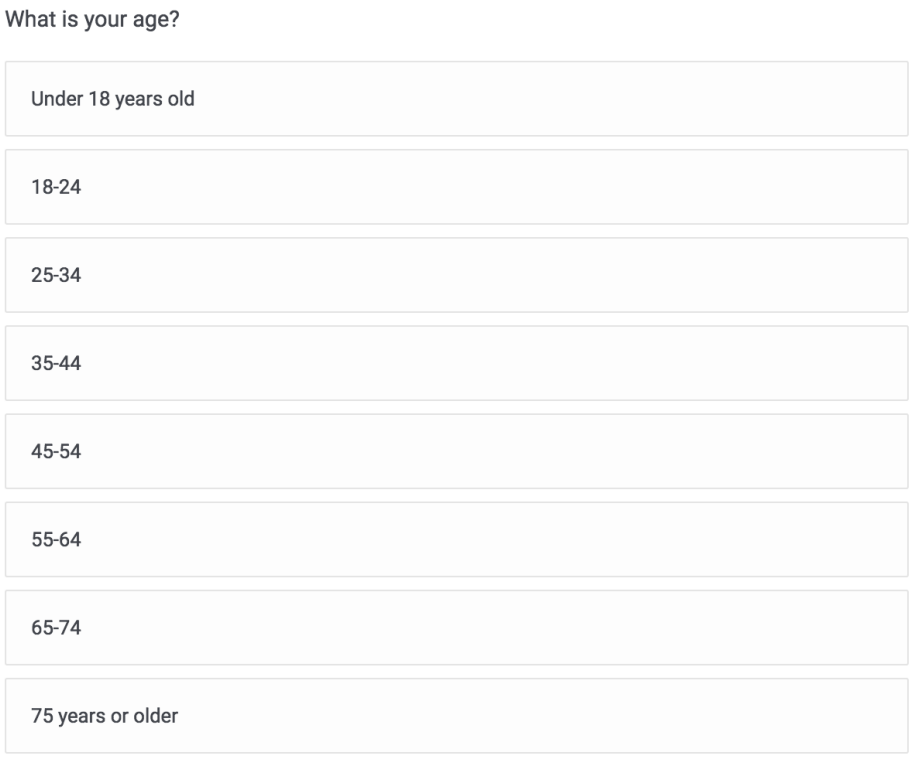

Figure B.1: Age

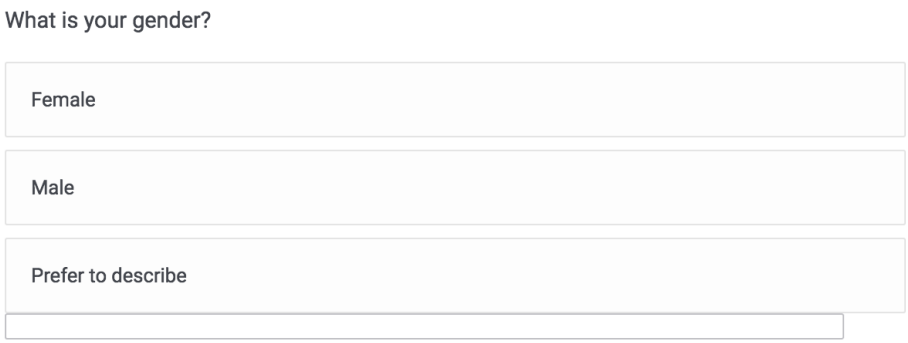

Figure B.2: Gender

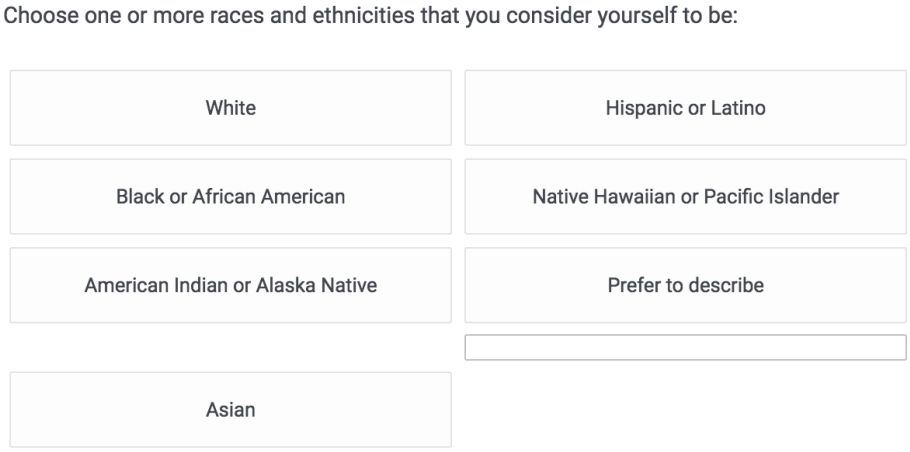

Figure B.3: Race

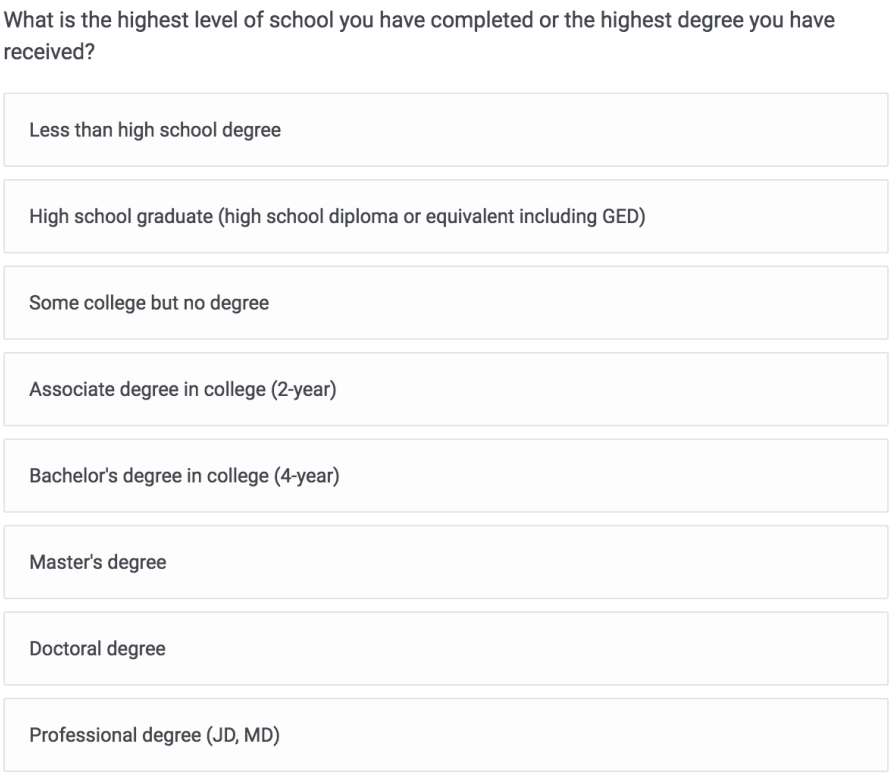

Figure B.4: Education Level

C. Catch-Trial Questions

These questions appeared at random points in the study to check if a participant was paying attention to the task. Each question has an easily identifiable correct answer. If a participant answered any of these questions incorrectly, their response was not used in the study.

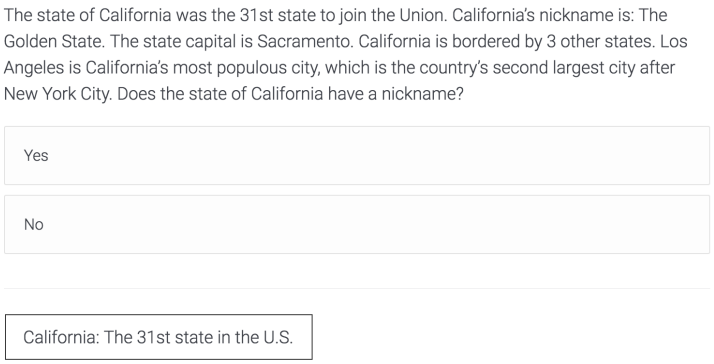

Figure C.1: Question to check if participant was paying attention. Correct answer is "Yes".

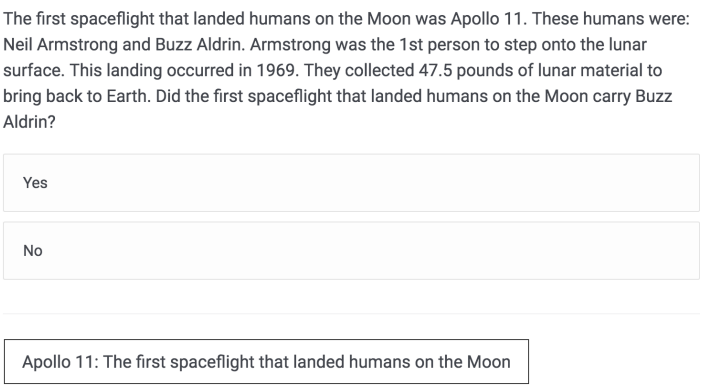

Figure C.2: Question to check if participant was paying attention. Correct answer is "Yes".

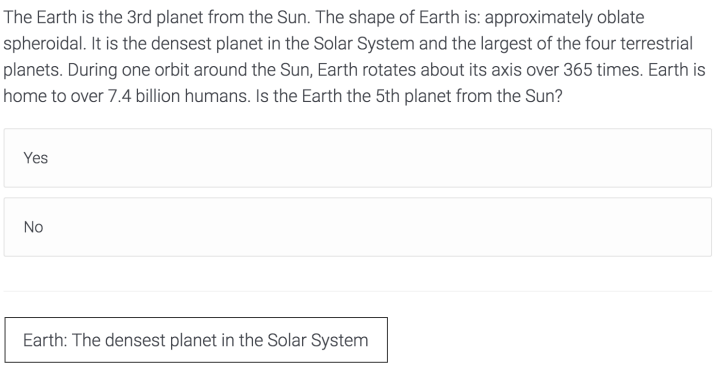

Figure C.3: Question to check if participant was paying attention. Correct answer is "No".

D. Results by Demographic

The results of the participants in the first study were analyzed by their various demographics to determine whether one's identity affected their accuracy in predicting recidivism.

Of the 400 participants, 214 self-identified as male and 185 self-identified as female. One participant self-identified as agender, so they were not included in either demographic. The average accuracy of the male participants was 62.1%, and the average accuracy of the female participants was also 62.5%. These accuracies were compared through a matched-pairs t-test of the 20 independent blocks. The test statistic was not significant at the 0.05 critical alpha level, t(19)= -0.5327, p= 0.6004. Therefore, there is not sufficient evidence to suggest that the difference in accuracy between the male participants and the female participants is significantly different.

Of the 400 participants, 173 reported to be 34 years old or younger, and 227 reported to be over 34 years old. The average accuracy of the younger participants was 62.4%, and the average accuracy of the older participants was 62.2%. The test statistic for a matched-pairs t-test was not significant at the 0.05 critical alpha level, t(19)= 0.4232, p= 0.6769. Therefore, there is not sufficient evidence to suggest that the difference in accuracy of the younger participants and the older participants is significantly different.

Of the 400 participants, 184 reported to have a Bachelors Degree or higher (Masters Degree, Doctoral Degree, or Professional Degree), and 216 reported to have an education level below a Bachelors Degree. The average accuracy of the participants with the higher level of education was 62.2%, and the average accuracy of the participants with the lower level of education was 62.4%. The test statistic for a matched-pairs t-test was not significant at the 0.05 critical alpha level, $t(19) = -0.2836$, $p = 0.7798$. Therefore, there is not sufficient evidence to suggest that the difference in accuracy of the more educated participants and the less educated participants is significantly different.

When asked about their race and ethnicity, 309 of the 400 participants identified as white. Because there were so few non-white participants, the accuracies of various races were not examined.

The results of this analysis suggest that ones demographic did not affect their accuracy in predicting recidivism.

E. Performance Comparison Between Studies by Defendant Race

The individual and crowd performances on the black and white defendants from each study were directly compared using matched-pairs t-tests of the 20 blocks to determine if the absence of race affected the participants' ability to predict recidivism for either race. Shown in Table E.1 are the results of these tests for the individual participant responses. Shown in Table E.2 are the results of these tests for the crowd performance. No values were significantly different between the studies.

Category	Without Race (%)	With Race (%)	t(19)	p-value
Black Accuracy	63.8	62.7	-1.6519	0.1150
Black False Positive	43.9	46.7	1.8367	0.0819
Black False Negative	31.5	29.8	-1.5345	0.1414
White Accuracy	63.8	65.0	1.5290	0.1427
White False Positive	33.8	30.9	-1.5815	0.1303
White False Negative	40.0	40.3	0.2631	0.7953

Table E.1: Results of t-tests comparing race performance.

Category	Without Race (%)	With Race(%)	t(19)	p-value
Black Accuracy	68.2	66.2	-1.7955	0.0885
Black False Positive	37.1	40.0	1.3580	0.1904
Black False Negative	29.2	30.1	0.6111	0.5484
White Accuracy	67.6	67.6	-0.0147	0.9884
White False Positive	27.2	26.2	-0.4383	0.6661
White False Negative	40.2	42.1	0.7080	0.4875

Table E.2: Results of t-tests comparing race performance for crowd.

References

[1] New data: Pretrial risk assessment tool works to reduce crime, increase court appearances. *Laura and John Arnold Foundation*, 2016.

[2] equivant Customer FAQ, 2017.

[3] Julia Angwin. Making Algorithms Accountable. *ProPublica*, 2016.

[4] Julia Angwin and Jeff Larson. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. *ProPublica*, 2016.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. *ProPublica*, 2016.

[6] Ben Popper. How Spotify's Discover Weekly cracked human curation at internet scale. *The Verge*, 2016.

[7] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

[8] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[9] Julia Carpenter. Google's algorithm shows prestigious job ads to men, but not to women. *The Independent*, 2015.

[10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. 2016.

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*, 2016.

[12] Penny Crosman. Can AI Be Programmed to Make Fair Lending Decisions? *American Banker*, 2016.

[13] Elizabeth R. Delong and North Carolina. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 2016.

[14] Hannah Delvin. Discrimination by algorithm: scientists devise test to detect AI bias. *The Guardian*, 2016.

[15] William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, July 2016.

[16] Richard O. Duda, Peter E. Hart, David G. Stork, C R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification, 2nd ed, 2001.

[17] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bais: There's Software used Across the Country to Predict Future Criminals. And it's Biased Against Blacks". *Federal Probation*, 80(2):38–46, 2014.

[18] William M. Grove, David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):19–30, 2000.

[19] James A Hanley and Barbara J McNeil. The Meaning and Use of the Area under a Receiver Operating (ROC) Curvel Characteristic. *Radiology*, 143(1):29–36, 1982.

[20] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. 2016.

[21] Reid Hastie and Tatsuya Kameda. The robust beauty of majority rules in group decisions. *Psychological Review*, 112(2):494–508, 2005.

[22] Vivian Ho. Seeking a better bail system, SF turns to computer algorithm. *San Francisco Chronicle*, 2016.

[23] Richard E. Hornsby. Florida Crimes and Offenses, 2017.

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. 2016.

[25] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, 2016.

[26] Adam Liptak. Sent to Prison by a Software Program's Secret Algorithms. *The New York Times*, 2017.

[27] Joshua A Markman, Matthew R Durose, Ramona R Rantala, and Andrew D Tiedt. Recidivism of Offenders Placed on Federl Community Supervision in 2005: Patterns from 2005 to 2010. Technical report, June 2016.

[28] P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

[29] Cecilia Muñoz, Megan Smith, and DJ Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Technical report, May 2016.

[30] Northpointe. Practitioner's Guide to COMPAS Core. Technical report, March 2015.

[31] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA, 2016.

[32] Cynthia Rudin. New models to predict recidivism could provide better way to deter repeat crime. *The Conversation*, 2015.

[33] Ben Sullivan. A New Program Judges If You're a Criminal From Your Facial Features. *Motherboard*, 2016.

[34] J. Surowiecki. *The Wisdom of Crowds*. Knopf Doubleday Publishing Group, 2005.

[35] Revati Thatte. UC Berkeley study dispels race as major factor in predicting future offenders. *The Daily Californian*, 2016.

[36] J. Stephen Wormith and Colin S. Goldstone. The clinical and statistical prediction of recidivism. *Criminal Justice and Behavior*, 11(1):3–34, 1984.