Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments

Ben Green
Harvard University
bgreen@g.harvard.edu

Yiling Chen
Harvard University
yiling@seas.harvard.edu

ABSTRACT

Despite vigorous debates about the technical characteristics of risk assessments being deployed in the U.S. criminal justice system, remarkably little research has studied how these tools affect actual decision-making processes. After all, risk assessments do not make definitive decisions—they inform judges, who are the final arbiters. It is therefore essential that considerations of risk assessments be informed by rigorous studies of how judges actually interpret and use them. This paper takes a first step toward such research on human interactions with risk assessments through a controlled experimental study on Amazon Mechanical Turk. We found several behaviors that call into question the supposed efficacy and fairness of risk assessments: our study participants 1) underperformed the risk assessment even when presented with its predictions, 2) could not effectively evaluate the accuracy of their own or the risk assessment's predictions, and 3) exhibited behaviors fraught with "disparate interactions," whereby the use of risk assessments led to higher risk predictions about black defendants and lower risk predictions about white defendants. These results suggest the need for a new "algorithm-in-the-loop" framework that places machine learning decision-making aids into the sociotechnical context of improving human decisions rather than the technical context of generating the best prediction in the abstract. If risk assessments are to be used at all, they must be grounded in rigorous evaluations of their real-world impacts instead of in their theoretical potential.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → *Law*.

KEYWORDS

fairness, risk assessment, behavioral experiment, Mechanical Turk

ACM Reference Format:
Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287563

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01.
https://doi.org/10.1145/3287560.3287563

1 INTRODUCTION

Across the United States, courts are increasingly using risk assessments to estimate the likelihood that criminal defendants will engage in unlawful behavior in the future.[1] These tools are being deployed during several stages of criminal justice adjudication, including at bail hearings (to predict the risk that the defendant, if released, will be rearrested before trial or not appear for trial) and at sentencing (to predict the risk that the defendant will recidivate). Because risk assessments rely on data and a standardized process, many proponents believe that they can mitigate judicial biases and make "objective" decisions about defendants [9, 12, 34]. Risk assessments have therefore gained widespread support as a tool to reduce incarceration rates and spur criminal justice reform [9, 27, 34].

Yet many are concerned that risk assessments make biased decisions due to the historical discrimination embedded in training data. For example, the widely-used COMPAS risk assessment tool wrongly labels black defendants as future criminals at twice the rate it does for white defendants [3]. Prompted by these concerns, machine learning researchers have developed a rapidly-growing body of technical work focused on topics such as characterizing the incompatibility of different fairness metrics [6, 44] and developing new algorithms to reduce bias [24, 33].

Despite these efforts, current research into fair machine learning fails to capture an essential aspect of how risk assessments impact the criminal justice system: their influence on judges. After all, risk assessments do not make definitive decisions about pretrial release and sentencing—they merely aid judges, who must decide whom to release before trial and how to sentence defendants after trial. In other words, algorithmic outputs act as decision-making aids rather than final arbiters. Thus, whether a risk assessment *itself* is accurate and fair is of only indirect concern—the primary considerations are how it affects decision-making processes and whether it makes *judges* more accurate and fair. No matter how well we characterize the technical specifications of risk assessments, we will not fully understand their impacts unless we also study how judges interpret and use them.

This study sheds new light on how risk assessments influence human decisions in the context of criminal justice adjudication. We ran experiments using Amazon Mechanical Turk to study how people make predictions about risk, both with and without the aid of a risk assessment. We focus on pretrial release, which in many respects resembles a typical prediction problem.[2] By studying

[1] Although there have been several generations of criminal justice risk assessments over the past century [41], throughout this paper we use risk assessments to refer to machine learning algorithms that provide statistical predictions.

[2] After someone is arrested, courts must decide whether to release that person until their trial. This is typically done by setting an amount of "bail," or money that the defendant must pay as collateral for release. The broad goal of this process is to protect individual liberty while also ensuring that the defendant appears in court for trial

behavior in this controlled environment, we discerned important patterns in how risk assessments influence human judgments of risk. Although these experiments involved laypeople rather than judges—limiting the extent to which our results can be assumed to directly implicate real-world risk assessments—they highlight several types of interactions that should be studied further before risk assessments can be responsibly deployed in the courtroom.

Our results suggest several ways in which the interactions between people and risk assessments can generate errors and biases in the administration of criminal justice, thus calling into question the supposed efficacy and fairness of risk assessments. First, even when presented with the risk assessment's predictions, participants made decisions that were less accurate than the advice provided. Second, people could not effectively evaluate the accuracy of their own or the risk assessment's predictions: participants' confidence in their performance was *negatively* associated with their actual performance and their judgments of the risk assessment's accuracy and fairness had no association with the risk assessment's actual accuracy and fairness. Finally, participant interactions with the risk assessment introduced two new forms of bias (which we collectively term "disparate interactions") into decision-making: when evaluating black defendants, participants were 25.9% more strongly influenced to increase their risk prediction at the suggestion of the risk assessment and were 36.4% more likely to deviate from the risk assessment toward higher levels of risk. Further research is necessary to ascertain whether judges exhibit similar behaviors.

The chain from algorithm to person to decision has become vitally important as algorithms inform increasing numbers of high-stakes decisions. To improve our understanding of these contexts, we introduce an "algorithm-in-the-loop" framework that places algorithms in a sociotechnical context—thus focusing attention on human-algorithm interactions to improve human decisions rather than focusing on the algorithm to improve its decisions. Rigorous studies of algorithm-in-the-loop systems are necessary to inform the design and implementation of algorithmic decision-making aids being deployed in the criminal justice system and beyond.

2 RELATED WORK

Despite some indications that risk assessments impact judges' decisions [29, 43, 58], little is known about the specific ways in which they influence judges. The most extensive study of this topic evaluated the changes prompted by Kentucky mandating in 2011 that risk assessments be used to inform all pretrial release decisions [59]. Although the risk assessment recommended immediate non-financial release for 90% of defendants, in practice the non-financial release rate increased only marginally (to 35%) before declining back toward the original release rate. The analysis found that the risk assessments had no effect on racial disparities. Two sets of related work provide further hints regarding how judges might use or otherwise respond to the predictions made by risk assessments.

and does not commit any crimes while released (whether the defendant is guilty of the offense that led to the arrest is not a factor at this stage). In order to make pretrial release decisions, judges must determine the likelihood—or the "risk"—that the defendant, if released, will fail to appear in court or will be arrested.

2.1 People are bad at incorporating quantitative predictions

The phenomenon of "automation bias" suggests that automated tools influence human decisions in significant, and often detrimental, ways. Two types of errors are particularly common: omission errors, in which people do not recognize when automated systems err, and commission errors, in which people follow automated systems without considering contradictory information [51]. Heavy reliance on automated systems can alter people's relationship to a task by creating a "moral buffer" between their decisions and the impacts of those decisions [11]. Thus, although "[a]utomated decision support tools are designed to improve decision effectiveness and reduce human error, [...] they can cause operators to relinquish a sense of responsibility and subsequently accountability because of a perception that the automation is in charge" [11].

Even when algorithms are more accurate, people do not appropriately incorporate algorithmic recommendations to improve their decisions, instead preferring to rely on their own or other people's judgment [47, 66]. One study found that people could not distinguish between reliable and unreliable predictions [30], and another found that people often deviate incorrectly from algorithmic forecasts [18]. Compounding this bias is the phenomenon of "algorithm aversion," through which people are less tolerant of errors made by algorithms than errors made by other people [17].

2.2 Information filters through existing biases

Previous research suggests that information presumed to help people make fairer decisions can fail to do so because it filters through people's preexisting biases. For example, "ban-the-box" policies (which are intended to promote racial equity in hiring by preventing employers from asking job applicants whether they have a criminal record) actually increase racial discrimination by allowing employers to rely on stereotypes and thereby overestimate how many black applicants have criminal records [2, 19]. Similarly, people's interpretations of police-worn body camera footage are significantly influenced by their prior attitudes about police [57].

Studies have shown that judges harbor implicit biases and that racial disparities in incarceration rates are due in part to differential judicial decisions across race [1, 54]. In Florida, for example, white judges give harsher sentences to black defendants than white ones who have committed the same crime and received the same score from the formula the state uses to set criminal punishments [55].

3 STUDY DESIGN

We conducted this study in two stages: first, developing a risk assessment for a population of criminal defendants, and second, running experiments on Mechanical Turk to determine how people incorporate these assessments into their own predictions.[3]

Before running our experiments, we made three hypotheses:

Hypothesis 1 (Performance). Participants presented with a risk assessment will make predictions that are less accurate than the risk assessment's.

[3]This study was reviewed and approved by the Harvard University Area Institutional Review Board and the National Archive of Criminal Justice Data.

Hypothesis 2 (Evaluation). Participants will be unable to accurately evaluate their own and the algorithm's performance.

Hypothesis 3 (Bias). As they interact with the risk assessment, participants will be disproportionately likely to increase risk predictions about black defendants and to decrease risk predictions about white defendants.

3.1 Defendant population and risk assessment

Stage 1 of the study involved developing a risk assessment for criminal defendants being considered for pretrial release (to predict the likelihood that, if released, they would be arrested before trial or fail to appear in court for trial). The goal of this stage was not to develop an optimal pretrial risk assessment, but to develop a risk assessment that resembles those used in practice and that could be presented to participants during the experiments in Stage 2.

We used a dataset collected by the U.S. Department of Justice that contains court processing information about 151,461 felony defendants who were arrested between 1990 and 2009 in 40 of the 75 most populous counties in the U.S. [61]. We restricted our analysis to defendants whose race was recorded as either black or white and who were released before trial, thus providing us with ground truth data about outcomes for each defendant. This yielded a dataset of 47,141 defendants (Table A.1). We pooled together failing to appear and being rearrested, defining any incidence of one or both of these outcomes as violating the terms of pretrial release; 29.8% of released defendants committed a violation.

After splitting the data into train and test sets, we trained a model (i.e., the risk assessment) using gradient boosted trees [26]. The model was based on five features about each defendant: age, offense type, previous failures to appear, and number of prior arrests and convictions. We excluded race and gender from the model to follow common practice among risk assessment developers [46]. Because our experiment participants would be predicting risk in increments of 10% (see Section 3.2), we rounded each risk assessment prediction to the nearest 10%.

The model achieves an area under the curve (AUC) of 0.67 on the test set, indicating comparable accuracy to COMPAS [37, 45], the Public Safety Assessment [13], and other risk assessments [14, 15]. We also evaluated the risk assessment model for fairness and found that it is well-calibrated (Figure A.1). We focused on calibration not as an ideal metric for fairness (recognizing that no perfect metric for fairness can exist [32]), but because it is the most commonly-used approach for evaluating risk assessments in practice [16, 25, 44]. Based on these attributes, our risk assessment resembles those used within U.S. courts.

We selected from the test set an experimental sample of 500 defendants whose profiles would be presented to both the control and treatment groups during the experiments (Table A.1).

The full details of how we developed the risk assessment and selected the sample population are available in the Appendix.

3.2 Experimental setup

In Stage 2 of the study, we conducted behavioral experiments on Amazon Mechanical Turk to determine how people use and are influenced by machine learning algorithms when making predictions about pretrial release. Each trial consisted of a consent page, a tutorial (with a description of the task and background information about pretrial release), an intro survey (Figure A.2), a series of predictions (described below), and an exit survey (Figure A.3). Both the intro and exit surveys included a simple question designed to ensure participants were paying attention. We also included a comprehension test with several multiple choice questions at the end of the tutorial; participants were not allowed to participate in the experiment until they correctly answered all of these questions. We restricted the task to Mechanical Turk workers who had an historical acceptance rate of $\geq 75\%$ and were inside the U.S. Each worker was allowed to participate in the experiment only once.

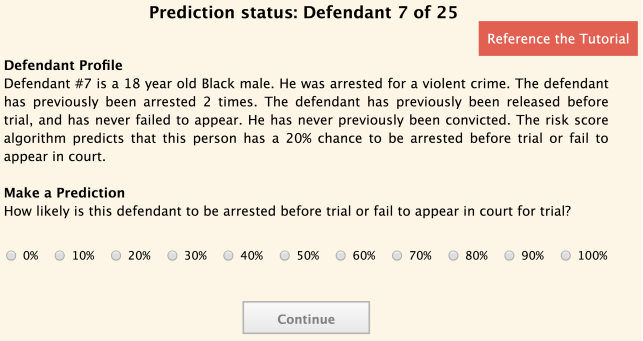

Figure 1: An example of the prompt presented to participants in the treatment group. Participants in the control group saw the same prompt, but without the sentence about the risk score algorithm.

The prediction task required participants to assess the likelihood that criminal defendants who have been arrested will commit a crime or fail to appear in court if they are released before trial (on a scale from 0% to 100%, in intervals of 10%). Each participant was presented with narrative profiles about a random sample of 25 defendants drawn from the 500-person experiment sample population. These profiles included the five features that the risk assessment incorporated as well as the race and gender of each defendant (we included these latter two features in the profiles because judges are exposed to these attributes in practice). While making predictions, participants could reference the tutorial to look up background information about pretrial release and the definitions of key terms.

When participants entered the experiment, they were randomly sorted into a control or treatment group; participants in the control group were shown the demographic information for each defendant, while participants in the treatment group were shown the risk assessment's prediction in addition to demographic information (Figure 1). We presented the same set of 500 defendants to both the control and treatment groups, allowing us to directly measure the impact on predictions of showing a risk assessment.

Participants were paid a base sum of $2 for completing the survey, with the opportunity to gain an additional reward of up to $2 based on their performance during the experiment. We allocated rewards according to a Brier score function, mapping the Brier reward (bounded [0,1], see Section 4.1) for each prediction to a payment using the formula $payment = reward * \$0.08$ (since the test population is restricted to defendants who were released

before trial, we have ground truth data with which to evaluate each prediction). Because the Brier score is a proper score function [28], participants were incentivized to report their true estimates of crime risk. We explicitly articulated this to participants during the tutorial and included a question about the reward structure in the comprehension test to ensure that they understood.

4 RESULTS

We conducted trials on Mechanical Turk over the course of a week in June 2018 (in 6 batches over 4 weekdays and 2 weekend days, at times ranging from morning to evening to account for variations in the population of Turk workers). 601 workers completed the experiment; we excluded all data from participants who failed at least one of the attention check questions or who required more than three attempts to pass the comprehension test. This process yielded a population of 554 participants (Table A.2). The participants were 58.5% male and 80.5% white, and the majority (65.5%) have completed at least a college degree. We asked participants to self-report their familiarity with machine learning and the U.S. criminal justice system on a scale from 1 ("Not at all") to 5 ("Extremely").

During the exit surveys, participants reported that the experiment paid well, was clear, and was enjoyable. Participants earned an average bonus of $1.54 (median=$1.56), making the average total payment $3.54. Participants completed the task in an average of 20 minutes (median=12), and earned an average wage of $20 per hour (median=$18). Out of 213 participants who responded to a free text question in the exit survey asking for any further comments, 32% mentioned that the experiment length and payment were fair. Participants were also asked in the exit survey to rate how clear and enjoyable the experiment was, on a scale from 1 to 5. The average rating for clarity was 4.4 (55% of participants rated the experiment clarity a 5), and the average rating for enjoyment was 3.6 (56% rated the experiment enjoyment a 4 or 5).

The participants cumulatively made 13,850 predictions about defendants, providing us with 13.85 ± 3.9 predictions about each defendant's risk under each of the two experimental conditions.

4.1 Analysis

We evaluated the accuracy and calibration of each prediction using the Brier reward: $reward = \left[1 - (prediction - outcome)^2\right]$, where $prediction \in \{0, 0.1, \ldots, 1\}$ and $outcome \in \{0, 1\}$ (thus, $reward \in [0, 1]$). When presented with a defendant who does not violate pretrial release, for example, a prediction of 0% risk would yield a reward of 1, a prediction of 100% would yield a reward of 0, and a prediction of 50% would yield a reward of 0.75. We also measured false positive rates (using a threshold of 50%).

Because we presented the same set of 500 defendants to both the control and treatment groups, we could measure the influence of the risk scores on the predictions about each defendant by comparing the predictions made by the control and treatment groups. For each defendant j, we defined the risk score's influence

$$I_j = \frac{t_j - c_j}{r_j - c_j} \tag{1}$$

where t_j and c_j are the average predictions made about that defendant by participants in the treatment and control groups, respectively, and r_j is the prediction made by the risk assessment. An $I = 0$ means that, on average, the treatment group makes identical predictions to the control group, completely discounting the risk score, while an $I = 1$ means that the treatment group makes identical predictions to the risk score.[4] This measure of influence is similar to the "weight of advice" metric that has been used to measure how much people alter their decisions when presented with advice [48, 65]. Comparing the distributions of predictions made by the control and treatment groups indicates that the risk assessment influences the full distribution of predictions made by the treatment group, not just the average (Figure A.4). To obtain reliable measurements, when evaluating algorithm influence we excluded all predictions about the 112 defendants for whom $|r_j - c_j| < 0.05$.

We used a variant of Equation 1 to measure the influence of the risk assessment on each participant in the treatment group. For every prediction made by a participant, we measured the risk assessment's influence by taking that prediction in place of the average treatment group prediction. We then averaged these influences across the 25 predictions that the participant made. That is, the influence of the risk assessment on participant k is

$$I^k = \frac{1}{25} \sum_{i=1}^{25} \frac{p_i^k - c_i}{r_i - c_i} \tag{2}$$

where p_i^k refers to participant k's prediction about the ith defendant (out of 25) presented.

Our primary dimension of analysis was to compare behavior and performance across the race of defendants, which has been at the crux of debates about fairness in criminal justice risk assessments [3, 6, 27]. Similar audits should be conducted across other intersecting forms of identity, such as gender and class [10].

4.2 Hypothesis 1 (Performance)

Participants in the treatment group earned a 4.0% larger average reward and a 16.4% lower false positive rate than participants in the control group (Table 1). A two-sided t-test and χ^2 test confirm that these differences are statistically significant (both with $p < 10^{-5}$). A regression of each participant's performance on their treatment and personal characteristics found that being in the treatment group was associated with a 0.03 higher average reward ($p < 10^{-7}$). The only personal attribute that had a significant relationship with average reward was gender (women performed slightly better than men, with $p = 0.045$).

Yet although presenting the risk assessment improved the performance of participants, the treatment group significantly underperformed the the risk assessment (Table 1). Despite being presented with the risk assessment's predictions, the treatment group achieved a 2.6% lower average reward and a 46.5% higher false positive rate than the risk assessment (both with $p < 10^{-8}$). Only 23.7% of participants in the treatment group earned a higher average reward than the risk assessment over the course of their trial, compared to 64.1% who earned a lower reward than the risk assessment (Figure A.5).

We broke these results down by race to compare how participants and the risk assessment performed when making predictions about black and white defendants. As Figure 2 indicates, a similar

[4] Although I will mostly fall between 0 and 1, it is possible for I to fall outside these bounds if participants move in the opposite direction than the risk assessment suggests or adjust beyond the risk assessment.

	Control N=6,250	Treatment N=7,600	Risk assessment N=7,600
Average reward	0.756	0.786	0.807
False positive rate	17.7%	14.8%	10.1%

Table 1: The first two columns show the performance of participants within the control and treatment groups and the third column shows the performance of the risk assessment (N is the total number of predictions made). Two-sided t-tests and χ^2 tests confirm that the average rewards and the false positive rates, respectively, of all three prediction approaches are statistically distinct from one another (all with $p < 10^{-5}$).

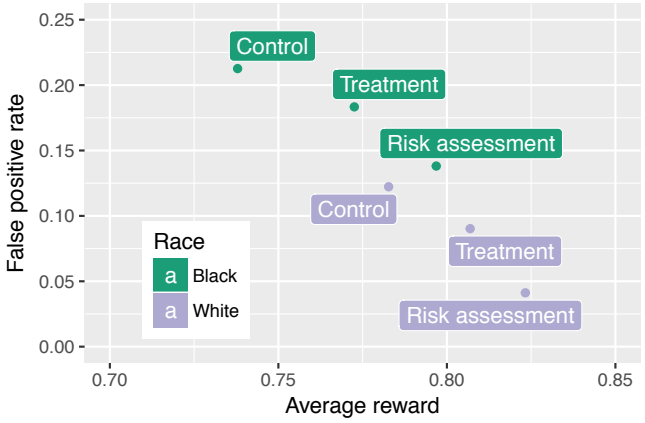

Figure 2: Performance of the control group, treatment group, and risk assessment, broken down by defendant race. In both cases, the treatment group outperforms the control group but underperforms the risk assessment.

pattern was true for both races: the treatment group outperformed the control group but underperformed the risk assessment. Taking the control group performance as a lower bound and the risk assessment performance as an upper bound, the treatment group achieved a similar relative improvement in its predictions about both races: for average reward, 58.7% of possible improvement for black defendants and 59.7% for white defendants; for false positive rate, 39.3% of possible improvement for black defendants and 39.5% for white defendants (neither difference across race is statistically significant).

The actual performance level differs significantly across race, however. All three prediction approaches (i.e., the control group, the treatment group, and the risk assessment) achieve a larger reward and lower false positive rate for white defendants than for black defendants (all with $p < 10^{-6}$). Most notably, the treatment group attains a 4.5% higher average reward for white than black defendants and its false positive rate for black defendants (18.3%) is more than double its false positive rate for white defendants (9.0%).

4.3 Hypothesis 2 (Evaluation)

To assess whether participants could evaluate the quality of their predictions, we compared their self-reported confidence (from the exit survey) to their actual performance, as measured by their average Brier reward during the task. The average participant confidence was 3.2 (on a scale from 1 to 5), with the reward decreasing as reported confidence increases (Figure A.6). We regressed confidence on performance (controlling for each participant's treatment, demographic information, and exit survey responses) and found that average reward was negatively associated with confidence ($p = 0.0186$). In other words, the more confidence participants expressed in their predictions, the less well they actually performed. This pattern holds across both the control and treatment groups.

We next analyzed whether participants in the treatment group could evaluate the risk assessment's accuracy, as measured by its average Brier reward on the 25 defendants presented to the participant (these average rewards ranged from 0.69 to 0.91). We regressed the participants' evaluations of the risk assessment's accuracy against the risk assessment's actual performance, while controlling for each participant's performance, demographic information, and exit survey responses. The participant's evaluation of the risk assessment's accuracy did not have any significant relationship with the risk assessment's performance during the task, suggesting that participants were unable to perceive any differences in risk assessment accuracy over the samples they observed (Figure A.6).

We also considered whether participants could discern how fairly the risk assessment made predictions. As a rough measure of algorithmic fairness during each trial, we measured the difference between the risk assessment's false positive rates for black and white defendants on the 25 defendants presented to the participant (in order to focus on the most salient aspect of bias, we restricted this analysis to the 81% of participants for whom the risk assessment had a greater or equal false positive rate for black than white defendants). Regressing participant evaluations of the risk assessment's fairness on the risk assessment's false positive rate differences (controlling for each participant's performance, demographic information, and exit survey responses, along with the risk assessment's performance) found no significant relationship between perceived and actual fairness (Figure A.6).

Finally, we evaluated whether participants in the treatment group could recognize how heavily they incorporated the risk assessment into their decisions. Regressing the participants' self-reports of influence on the extent to which they were actually influenced by the risk assessment (using the risk score influence measure introduced in Equation 2, and controlling for each participant's performance, demographic information, and exit survey responses, along with the risk assessment's performance) indicates that participants could generally discern how strongly they were influenced by the risk assessment ($p < 10^{-4}$; Figure A.6).

4.4 Hypothesis 3 (Bias)

We interrogated Hypothesis 3 through two complementary approaches: first, by taking the control group's predictions as the baseline participant predictions to measure the risk assessment's influence on the treatment group, and second, by taking the risk assessment's predictions as the starting point to measure how much

and in which direction the treatment group participants deviated from those predictions.

Although we could not precisely discern how participants made decisions, the responses to an optional free response question in the exit survey about how participants used the risk scores (Question 5 in Figure A.3) suggest that people predominantly followed a mix of these two approaches. Out of the 156 participants who described their strategy, 79 (50.6%) used the risk assessment as a baseline, 58 (37.2%) made their own judgment and then incorporated the risk assessment, 10 (6.4%) followed the risk assessment completely, and 9 (5.8%) ignored the risk assessment entirely (Table A.3). The group that followed the risk assessment earned the largest average reward (0.81), while the group that ignored the risk assessment earned the lowest (0.77). The other two groups both earned average rewards of 0.79, and were statistically indistinguishable.

Analyzing behavior through the lens of the two most common strategies yields complementary evidence for "disparate interactions," i.e., interactions with the risk assessment that lead participants to disproportionately make higher risk predictions about black defendants and lower risk predictions about white defendants.

4.4.1 Influence of risk scores. Because we presented the same population of defendants to the control and treatment groups, we could directly measure how presenting the risk score to participants affected the predictions made about each defendant. For each defendant, we measured the influence of the risk assessment on the treatment group's predictions as described in Equation 1 (excluding the 112 defendants for whom $|r_j - c_j| < 0.05$). The risk assessment exhibited an average influence of 0.61; as this number is greater than 0.5, it suggests that treatment group participants placed more weight on the risk assessment than on their own judgment. A two-sided t-test found no statistically significant difference between the risk assessment's influence when its prediction was less or greater than the control group's prediction ($r < c$ or $r > c$, respectively).

Splitting the defendants by race tells a more complex story (Figure 3). When the risk score was lower than the control group's average prediction ($r < c$), the risk assessment exerted a similar influence on participants regardless of the defendant's race (0.61 vs. 0.60; p=0.77). Yet when the risk assessment predicted a higher risk than the control group ($r > c$), it exerted a 25.9% stronger average influence on predictions about black defendants than on predictions about white defendants (0.68 vs. 0.54; a two-sided t-test finds $p = 0.02$ and 95CI of the difference in means [0.02, 0.25]).

This outcome cannot be explained by differences in the raw disparities between the risk assessment's and the control group's predictions (i.e., the value of $r - c$), since the values of $r - c$ do not differ significantly across defendant race (the average disparity for both races is 0.25 when $r < c$ and 0.11 when $r > c$). Breaking out Figure 3 based on the value of $r - c$ indicates that the risk assessment exerts an equal influence on predictions about both races at all values of $r - c$, except for when $r - c = 0.1$ (Figure A.7).

Thus, the risk assessment leads to larger increases in risk for black defendants (as measured by $t-c$). While the shift in participant predictions precipitated by the risk assessment is identical when $r < c$ (the risk assessment generates an average reduction of 0.14 for both black and white defendants), when $r > c$ the average increase for black defendants is 0.075 while the average increase for white

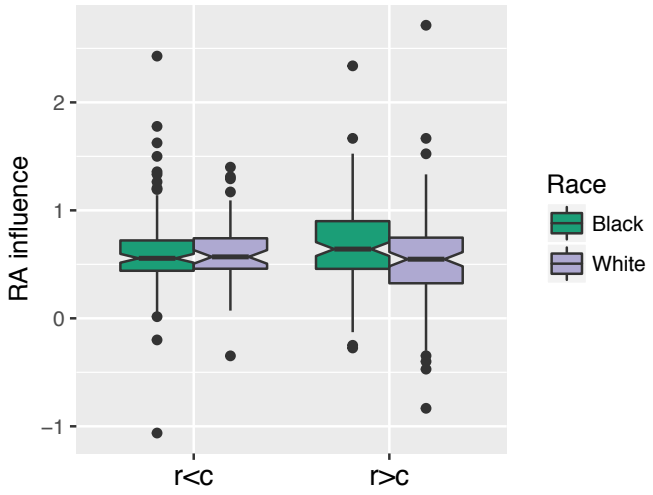

Figure 3: The influence of the risk assessment (RA) on participant predictions, broken down by whether the risk score is less or greater than the control group's average prediction (r < c and r > c, respectively), and compared across the race of defendants. While the risk assessment's influence is nearly identical across race when r < c, when r > c the risk assessment exerts a 25.9% stronger influence on participants who are evaluating black defendants (p = 0.02).

defendants is 0.063. Although these results are not significant (a two-sided t-test finds $p = 0.076$ and 95CI difference in means [-0.001, 0.02]), considering each prediction from the treatment group independently, rather than taking averages for each defendant (i.e., replacing t_j with p_j^k in Equation 1), yields further evidence for this result: the average increase for black defendants is 0.077 compared to 0.064 for white defendants (a 20.3% larger average increase), with $p = 0.003$ and 95CI difference in means [0.004, 0.02]. Moreover, among defendants for whom $r - c = 0.1$, the increase in participant risk prediction instigated by the risk assessment is 25.5% larger for black defendants ($p = 0.042$; Figure A.7).

We ran linear regressions to see what determines the risk assessment's influence on participants. We split defendants into two categories—those for whom $r < c$ (Group 1) and those for whom $r > c$ (Group 2). For each group, we regressed the algorithm's influence on predictions about each defendant (Equation 1) on that defendant's demographic attributes and criminal background, along with the value of $|r - c|$. For Group 1, the risk assessment exerted more influence as $|r - c|$ increased, but less influence for defendants with a previous failure to appear on their records. For Group 2, the risk assessment similarly was more influential as $|r - c|$ increased. Three other attributes were also statistically significant: the risk assessment exerted more influence on participants making predictions about black defendants, defendants who were arrested for a violent crime, and defendants with more prior convictions. Thus, when $r > c$, participants were more strongly influenced to increase their risk predictions for black defendants in two ways: they responded both directly to race and to a feature that is correlated with race (prior convictions; Table A.1).

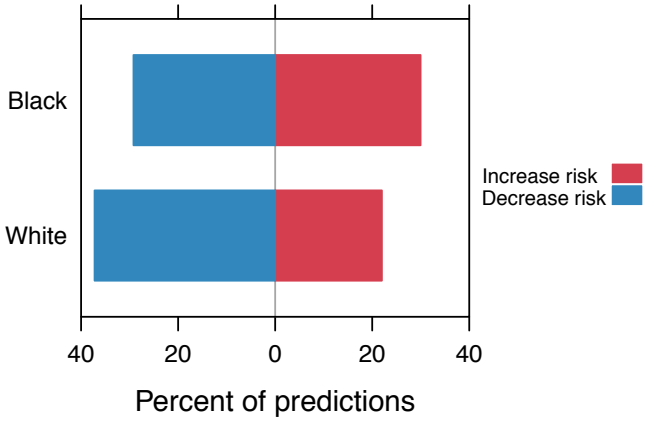

Figure 4: The rate at which participants deviated from the risk assessment's prediction toward higher and lower levels of risk, broken down by defendant race. When evaluating black defendants, participants were 36.4% more likely to deviate positively from the risk assessment and 21.5% less likely to deviate negatively (participant predictions matched the risk assessment at an equal rate for both races).

4.4.2 Participant deviations from risk scores. For each prediction made by participants in the treatment group, we measured how far and in which direction that prediction deviated from the risk assessment's recommendation. That is, we measured $d_j^k = p_j^k - r_j$. The average deviation among the 7600 treatment group predictions was 0.014, with a median deviation of 0. Participants deviated to a higher risk prediction 26.9% of the time, matched the risk assessment 40.8% of the time, and deviated to a lower risk prediction 32.3% of the time. The results from Section 4.2 suggest that these deviations tend to make participant predictions less accurate than the risk assessment.

As in the previous section, these statistics differ by defendant race. While the average deviation for white defendants was -0.002, the average deviation for black defendants was 0.024 ($p = 7 \times 10^{-13}$, 95CI difference in means [0.019, 0.033]). This difference emerged because participants were more likely to deviate positively from the risk assessment when evaluating black defendants and to deviate negatively when evaluating white defendants (the average deviation magnitude was the same across race for both positive and negative deviations). As Figure 4 depicts, participants deviated to a higher risk prediction 30.0% of the time for black defendants compared to 22.0% of the time for white defendants (36.4% more), and conversely deviated to a lower risk prediction 29.2% of the time for black defendants compared to 37.2% of the time for white defendants (21.5% less). Participants matched the risk assessment in 40.8% of predictions when evaluating both races.

We regressed each deviation on the characteristics of the defendant and the participant, the prediction made by the risk assessment, and the participant's status in the experiment (i.e., which in the sequence of 25 predictions the participant was making). Since these deviations include repeated samples for each defendant and participant, we used a linear mixed-effects model with random effects for the defendant and participant identities. Several characteristics of defendants had statistically significant associations with the deviations: participants were more likely to deviate positively from the risk assessment when evaluating younger defendants, defendants arrested for a violent crime, defendants with more prior arrests and convictions, and defendants with a prior failure to appear. Neither the defendant's race nor any attributes of participants had a statistically significant relationship with deviations.

These results suggest that while participants did not deviate from the risk assessment based explicitly on race, they deviated based on attributes that are unevenly distributed across race: compared to white defendants, black defendants on average have more prior arrests, convictions, and failures to appear (Table A.1).

5 DISCUSSION

This study presents initial evidence regarding how risk assessments influence human decision-makers. Confirming our three hypotheses, our results indicate that people underperform risk assessments even when provided with its advice; are unable to evaluate the performance of themselves or the risk assessment; and engage in "disparate interactions," whereby their use of risk assessments leads to higher risk predictions about black defendants and lower risk predictions about white defendants.

This work demonstrates how theoretical evaluations are necessary but insufficient to evaluate the impacts of risk assessments: what appears to be a fair source of information can, depending on how people interact with it, become a leverage point around which discrimination manifests. It is necessary to place risk assessments into a sociotechnical context so that their full impacts can be identified and evaluated.

Our results highlight a significant but often overlooked aspect of algorithmic decision-making aids: introducing risk assessments to the criminal justice system does not eliminate discretion to create "objective" judgments, as many have argued [9, 12, 34]. Instead, risk assessments merely shift discretion to different places, which include the judge's interpretation of the assessment and decision about how strongly to rely on it. This reality must become a central consideration of any proposals for and evaluations of risk assessments, especially given that previous attempts to standardize the criminal justice system—sentencing reform efforts in the 1980s—shifted discretion to prosecutors, generating a racially-biased rise in excessive punishment [49].

A particular danger of judicial discretion about how to incorporate risk assessments into decisions is the potential for disparate interactions: biases that emerge as an algorithmic prediction filters through a person into a decision. Our experiment participants were 25.9% more strongly influenced by the risk assessment to increase their risk prediction when evaluating black defendants than white ones, leading to a 20.3% larger average increase for black than white defendants due the risk assessment. Moreover, participants were 36.4% more likely to deviate positively from the risk assessment and 21.5% less likely to deviate negatively from the risk assessment when evaluating black defendants.[5]

[5] Although it is possible that participants predicted higher risk for black defendants to account for the racial bias in arrests, we do not believe this was an important factor since no participants mentioned any such thought process in the exit survey when describing their behavior.

These disparate interactions emerged through both direct and indirect bias: while race had a direct role in increasing the risk score's influence on participants, the disparities in influence and deviations also arose due to participants responding to particularly salient features that are unevenly distributed by race (such as number of prior convictions)—essentially double-counting features for which the risk assessment had already accounted. This behavior resembles that of machine learning algorithms, which can be racially biased even when race is not included as an explicit factor [3], and highlights the importance of studying the complex mechanisms through which discrimination can manifest. Future work should explore how different ways of presenting and explaining risk assessments (and of training people to use them) could improve performance and in particular reduce disparate interactions.

An important research direction that could guide such efforts is to study the processes through which people make decisions when provided with risk assessments. Our participants followed several approaches when evaluating defendants, the most common being using the risk assessment to influence their initial judgment and using the risk assessment as a baseline (Table A.3). Analyzing participant behavior from both of these perspectives indicated related forms of disparate interactions. Meanwhile, the most successful strategy was to directly follow the risk assessment. While in theory it is possible for people to synthesize the risk assessment with their own judgment to make better decisions than either could alone, in practice we found no evidence that any strategy taken by participants leads them to outperform the risk assessment.

A major limitation to people's use of risk assessments is their inability to evaluate their own and the risk assessment's performance. Many proponents defend the deployment of risk assessments on the grounds that judges have the final say and can discern when to rely on the predictions provided [37, 46, 63]. But our results indicate that this is an unrealistic expectation: our participants' judgments about their own performance were *negatively* associated with their actual performance, and their evaluations of the risk assessment had no statistically significant relationship with its actual performance (other research has similarly shown that people struggle to detect algorithmic mistakes across a variety of conditions [53]). Given these results, it is no wonder that participants in the treatment group underperformed the risk assessment. How can we expect people to navigate the balance between their own judgment and a risk assessment's when they are unable to accurately assess their own or the algorithm's performance in the first place? Determining how to incorporate a risk assessment into one's own prediction is arguably a more challenging task that requires more expertise than merely making a prediction.

The results of this study raise one of the most important but rarely-discussed issues at the heart of debates about risk assessments: how *should* risk assessments be incorporated into existing practices? On the one hand, risk assessments alone achieve better performance than individuals (both with and without a risk assessment's aid) in terms of accuracy and false positive rates.[6] Yet there are many reasons to be wary of relying too heavily on risk assessments, including due process concerns, their embedding of discriminatory and punitive approaches to justice, and their potential to hinder more systemic criminal justice reforms [7, 31, 58]. Meanwhile, the current approach of presenting predictions to judges without sufficient guidelines or training comes with the issues of poor interpretation and disparate interactions.

The conflicts between these positions are apparent in how the Wisconsin Supreme Court severely circumscribed the role of risk assessments in its decision in State v. Loomis, regarding the use of COMPAS in sentencing. Despite defending the use of COMPAS on the grounds that it "has the potential to provide sentencing courts with more complete information," the Court also mandated that "risk scores may not be used: (1) to determine whether an offender is incarcerated; or (2) to determine the severity of the sentence" [63]. If COMPAS is not supposed to influence the sentence, there are few purposes that the "more complete information" it provides can serve—and few ways to ensure that it serves only those purposes. In that case, why show it at all?

5.1 An Algorithm-in-the-Loop Framework

As computational systems permeate everyday life and inform critical decisions, it is of paramount importance to study how algorithmic predictions impact human decision-making across a broad range of contexts. Risk assessments are just one of an emerging group of algorithms that are intended to inform people making decisions (other examples include predictions to help companies hire job applicants and to help doctors diagnose patients). Yet despite robust research into the technical properties of these algorithms, we have a limited understanding of their sociotechnical properties: most notably, whether and how they actually improve decision-making. To answer these questions, it is necessary to study algorithms following the notion of "technologies as social practice," which is grounded in the understanding that technologies "are constituted through and inseparable from the specifically situated practices of their use" [60].

A natural body of work from which to draw inspiration in studying human-algorithm collaborations is human-in-the-loop (HITL) systems. In settings such as social computing and active learning, computational systems rely on human labor (such as labeling photos and correcting errors) to overcome limitations and improve their performance. But where HITL processes privilege models and algorithms, utilizing people where necessary to improve computational performance, settings like pretrial release operate in reverse, using algorithms to improve human decisions.

This distinction suggests the need for an alternative framework: algorithm-in-the-loop (AITL) systems.[7] Instead of improving computation by using humans to handle algorithmic blind spots (such as analyzing unstructured data), AITL systems improve human decisions by using computation to handle cognitive blind spots (such as finding patterns in large, complex datasets). This framework centers human-algorithm interactions as the locus of study and

[6] This result assumes a comparison between a single individual and a risk assessment. This is in contrast to a recent study suggesting that humans are just as accurate as COMPAS: that result holds only when the predictions of humans are aggregated to create a "wisdom of the crowd" effect; in fact, that study similarly found COMPAS to be more accurate than individuals [21].

[7] Although previous studies have used the phrase "algorithm-in-the-loop," they have defined it in the context of simulation and modeling rather than in relation to human-in-the-loop computations and human-algorithm interactions [56, 64].

prioritizes the human's decision over the algorithm's as the most important outcome.

An algorithm-in-the-loop perspective can inform essential sociotechnical research into algorithms. Recent work related to interpretability provides one important direction where progress is already being made [20, 52, 53]. Future analysis should focus on how to develop and present algorithms so that people can most effectively and fairly incorporate them into their deliberative processes, with particular attention to improving evaluations of algorithm quality and reducing disparate interactions. This may involve altering the algorithm in unintuitive ways: previous research suggests that in certain situations a seemingly suboptimal algorithm actually leads to better outcomes when provided to people as advice [23].

It will also be important to study the efficacy of different mechanisms for combining human and algorithmic judgment across a variety of contexts. Most algorithm-in-the-loop settings involve simply presenting an algorithmic output to a human decision-maker, relying on the person to interpret and incorporate that information. Yet research within human-computer interaction and crowdsourcing suggests that alternative approaches could lead to a better synthesis of human and computer intelligence [8, 35, 39, 40]. Which mechanisms are most effective (and desirable from an ethical and procedural standpoint) will likely vary depending on the situation.

Finally, given that automation can induce a moral buffer [11], it is necessary to study how using algorithms affects people's sense of responsibility for their decisions. Given the all-too-common expressions from engineers that they do not bear responsibility for the social impacts of their technologies [36, 62], the potential for automation bias raises the unsettling specter of situations in which both the engineers developing algorithms and the people using them believe the other to be primarily responsible for the social outcomes. It is of vital importance to study whether algorithms create a moral buffer and to find ways to avoid such scenarios.

5.2 Limitations

Given that our experiments were conducted on a population of Mechanical Turk workers rather than actual judges in the courtroom, it is necessary to circumscribe the interpretation of these results. Judges have more expertise than laypeople at predicting pretrial risk and are generally given more information about the risk assessments in use. Interestingly, however, judges have been shown to release many high-risk defendants and detain many low-risk ones [43]. Judges may also be more reluctant to rely on risk assessments, believing that their own judgment is superior: previous research has shown that people with more expertise are less willing to take advice [38, 48, 50], and a recent survey found that less than 10% of judges believed that an actuarial assessment could outperform their own predictions of risk [5].

Our study also fails to capture the level of racial priming that could influence judges' use of risk assessments. While our experiment tells participants that a defendant is black or white, a judge would also see the defendant's name and physical appearance. Studies have shown that employers discriminate based on racially-indicative names [4] and that judges are harsher toward defendants with darker skin and more Afrocentric features [22, 42]. Thus, it is possible that the disparate interactions we observe in our experiments could be heightened in the courtroom, where race is more salient. Future research should study how people respond to risk assessments as racial priming increases.

The short length of each trial (25 predictions over approximately 20 minutes) means that we could not capture how the relationships between people and risk assessments evolve over extended periods of time. This is an important factor to consider when deploying algorithmic systems, especially given research demonstrating that the changes instigated by risk assessments are short-lived [59]. The immediate impacts of introducing algorithms into decision-making processes may not indicate the long-term implications of doing so. This is particularly true within the criminal justice system, where political incentives and manipulation can distort the use of risk assessments over time [31].

Thus, while this study hints at issues that may arise in the courtroom, it remains an open question how closely our results resemble the outcomes of real-world implementation. Further studies must be done, in both experimental and natural settings, before risk assessments can be seriously considered for broader deployment in the criminal justice system, if they are to be used at all.

ACKNOWLEDGMENTS

The authors thank three anonymous reviewers for their constructive feedback. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] David S. Abrams, Marianne Bertrand, and Sendhil Mullainathan. 2012. Do Judges Vary in Their Treatment of Race? *The Journal of Legal Studies* 41, 2 (2012), 347–383.
[2] Amanda Agan and Sonja Starr. 2017. Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment. *The Quarterly Journal of Economics* 133, 1 (2017), 191–235.
[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[4] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004), 991–1013.
[5] Steven L. Chanenson and Jordan M. Hyatt. 2016. The Use of Risk Assessment at Sentencing: Implications for Research and Policy. *Bureau of Justice Assistance* (2016).
[6] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.
[7] Danielle Keats Citron. 2007. Technological Due Process. *Washington University Law Review* 85 (2007), 1249.
[8] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 329–340.
[9] New Jersey Courts. 2017. One Year Criminal Justice Reform Report to the Governor and the Legislature. (2017). https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf
[10] Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *The University of Chicago Legal Forum* (1989), 139.
[11] Mary L. Cummings. 2006. Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies* (2006).
[12] Mona J.E. Danner, Marie VanNostrand, and Lisa M. Spruance. 2015. Risk-Based Pretrial Release Recommendation and Supervision Guidelines.

[13] Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. 2018. The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky. (2018).

[14] Sarah L. Desmarais, Kiersten L. Johnson, and Jay P. Singh. 2016. Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings. *Psychological Services* 13, 3 (2016), 206–222.

[15] Sarah L. Desmarais and Jay P. Singh. 2013. Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States. (2013).

[16] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe Inc.* (2016).

[17] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[18] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2016. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* (2016). https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2643

[19] Jennifer L. Doleac and Benjamin Hansen. 2017. The Unintended Consequences of 'Ban the Box': Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden. (2017).

[20] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).

[21] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018).

[22] Jennifer L. Eberhardt, Paul G. Davies, Valerie J. Purdie-Vaughns, and Sheri Lynn Johnson. 2006. Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes. *Psychological Science* 17, 5 (2006), 383–386.

[23] Avshalom Elmalech, David Sarne, Avi Rosenfeld, and Eden Shalom Erez. 2015. When Suboptimal Rules. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 1313–1319.

[24] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.

[25] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks". *Federal Probation* 80, 2 (2016), 38–46.

[26] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29, 2 (2001), 1189–1232.

[27] Gideon's Promise, The National Legal Aid and Defenders Association, The National Association for Public Defense, and The National Association of Criminal Defense Lawyers. 2017. Joint Statement in Support of the Use of Pretrial Risk Assessment Instruments. (2017). http://www.publicdefenders.us/files/Defenders%20Statement%20on%20Pretrial%20RAI%20May%202017.pdf

[28] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

[29] John S. Goldkamp and Michael R. Gottfredson. 1984. Judicial Decision Guidelines for Bail: The Philadelphia Experiment. *National Institute of Justice Research Report* (1984).

[30] Paul Goodwin and Robert Fildes. 1999. Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? *Journal of Behavioral Decision Making* 12, 1 (1999), 37–53.

[31] Ben Green. 2018. "Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[32] Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *Proceedings of the Machine Learning: The Debates Workshop*.

[33] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 3315–3323.

[34] Kamala Harris and Rand Paul. 2017. Pretrial Integrity and Safety Act of 2017. *115th Congress* (2017).

[35] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 159–166.

[36] Matthew Hutson. 2018. Artificial intelligence could identify gang crimes—and ignite an ethical firestorm. *Science* (2018). http://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

[37] Northpointe Inc. 2012. COMPAS Risk & Need Assessment System. (2012). http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf

[38] Jonas Jacobson, Jasmine Dobbs-Marsh, Varda Liberman, and Julia A. Minson. 2011. Predicting Civil Jury Verdicts: How Attorneys Use (and Misuse) a Second Opinion. *Journal of Empirical Legal Studies* 8 (2011), 99–119.

[39] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 4070–4073.

[40] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. 467–474.

[41] Danielle Kehl, Priscilla Guo, and Samuel Kessler. 2017. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. *Responsive Communities Initiative, Berkman Klein Center for Internet & Society* (2017).

[42] Ryan D. King and Brian D. Johnson. 2016. A Punishing Look: Skin Tone and Afrocentric Features in the Halls of Justice. *Amer. J. Sociology* 122, 1 (2016), 90–124.

[43] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293.

[44] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv:1609.05807* (2016).

[45] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[46] Laura and John Arnold Foundation. 2016. Public Safety Assessment: Risk Factors and Formula. (2016). http://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf

[47] Joa Sang Lim and Marcus O'Connor. 1995. Judgemental Adjustment of Initial Forecasts: Its Effectiveness and Biases. *Journal of Behavioral Decision Making* 8, 3 (1995), 149–168.

[48] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2018. Algorithm Appreciation: People Prefer Algorithmic To Human Judgment. (2018).

[49] Mona Lynch. 2016. *Hard Bargains: The Coercive Power of Drug Laws in Federal Court.* Russell Sage Foundation.

[50] Paul E. Meehl. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press.

[51] Kathleen L. Mosier, Melisa Dunbar, Lori McDonnell, Linda J. Skitka, Mark Burdick, and Bonnie Rosenblatt. 1998. Automation Bias and Errors: Are Teams Better than Individuals? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42, 3 (1998), 201–205.

[52] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1802.00682* (2018).

[53] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[54] Jeffrey J. Rachlinski, Sheri Lynn Johnson, Andrew J. Wistrich, and Chris Guthrie. 2008. Does Unconscious Racial Bias Affect Trial Judges. *Notre Dame Law Review* 84 (2008), 1195–1246.

[55] Josh Salman, Emily Le Coz, and Elizabeth Johnson. 2016. Florida's broken sentencing system. *Sarasota Herald-Tribune* (2016). http://projects.heraldtribune.com/bias/sentencing/

[56] Palash Sarkar, Jukka Kortela, Alexandre Boriouchkine, Elena Zattoni, and Sirkka-Liisa Jämsä-Jounela. 2017. Data-Reconciliation Based Fault-Tolerant Model Predictive Control for a Biomass Boiler. *Energies* 10, 2 (2017), 194.

[57] Roseanna Sommers. 2016. Will Putting Cameras on Police Reduce Polarization? *Yale Law Journal* 125, 5 (2016), 1304–1362.

[58] Sonja B. Starr. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66, 4 (2014), 803–872.

[59] Megan Stevenson. forthcoming. Assessing Risk Assessment in Action. *Minnesota Law Review* 103 (forthcoming).

[60] Lucy Suchman, Jeanette Blomberg, Julian E. Orr, and Randall Trigg. 1999. Reconstructing Technologies as Social Practice. *American Behavioral Scientist* 43, 3 (1999), 392–408.

[61] United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. 2014. State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties.

[62] Tom Upchurch. 2018. To work for society, data scientists need a hippocratic oath with teeth. *Wired* (2018). https://www.wired.co.uk/article/data-ai-ethics-hippocratic-oath-cathy-o-neil-weapons-of-math-destruction

[63] Wisconsin Supreme Court. 2016. *State v. Loomis.* 881 Wis. N.W.2d 749.

[64] Quanzhong Yan, Minghui Kao, and Michael Barrera. 2010. Algorithm-in-the-Loop with Plant Model Simulation, Reusable Test Suite in Production Codes Verification and Controller Hardware-in-the-Loop Bench Testing. *SAE Technical Paper* 0148-7191 (2010).

[65] Ilan Yaniv. 2004. Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes* 93, 1 (2004), 1–13.

[66] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2017. Making Sense of Recommendations. (2017).

A APPENDIX: METHODS AND RESULTS

Stage 1 of the study involved developing a risk assessment for criminal defendants being considered for pretrial release (to predict the likelihood that a criminal defendant, if released, would be arrested before trial or fail to appear in court for trial).

The goal of this stage was not to develop an optimal algorithm for the purpose of pretrial risk assessment, but to develop a risk assessment that could be presented to participants during the experiments in Stage 2. The primary benchmark for the algorithm was that it make predictions more accurately than humans; given that that algorithms are more accurate than humans across a wide variety of tasks, and in particular in the domain of pretrial risk assessment [43], this benchmark was essential for creating a realistic experimental environment. The results described in Section 4.2 clearly indicate that this criterion was satisfied.

We used a dataset collected by the U.S. Department of Justice that contains court processing information about 151,461 felony defendants who were arrested between 1990 and 2009 in 40 of the 75 most populous counties in the United States [61]. The data includes information about arrest charges, demographic characteristics, criminal history, pretrial release and detention, adjudication, and sentencing (each row follows a specific case against an individual defendant; it is possible that the same person appears multiple times across different cases, but the data did not indicate individual identities). We cleaned the dataset by removing all records with missing values, and restricted our analysis to defendants who were at least 18 years old and whose race was recorded as either black or white. We further restricted our analysis to defendants who were released before trial, and thus for whom we had ground truth data about whether that person was rearrested or failed to appear before trial.

This process yielded a dataset of 47,141 released defendants (Table A.1). This population was 76.7% male and 55.7% black, with an average age of 30.8 years. The most common offense type was drug crimes (36.9%), followed by property crimes (32.7%), violent crimes (20.4%), and public order crimes (10.0%). 63.4% of defendants had previously been arrested and 46.5% had previously been convicted. Of the 29,875 defendants who had previously been released before a trial, 39.6% had failed to appear at least once. After being released, 15.0% of defendants were rearrested before trial and 20.3% of defendants failed to appear for trial. We pooled these outcomes together and defined any incidence of one or both of these outcomes as violating the terms of pretrial release; 29.8% of released defendants committed a violation.

We randomly split the dataset into train and test sets with 80% and 20% of the records, respectively. We then trained a model using gradient boosted trees [26] to predict which defendants would violate pretrial release (i.e., which defendants would be rearrested or fail to appear in court), based on five features about each defendant: age, offense type, number of prior arrests, previous failure to appear, and number of prior convictions. We excluded race and gender from the model to match common practice among risk assessment developers [46]. Because our experiment participants would be predicting risk in increments of 10% (see Section 3.2), we rounded each prediction to the nearest 10%.

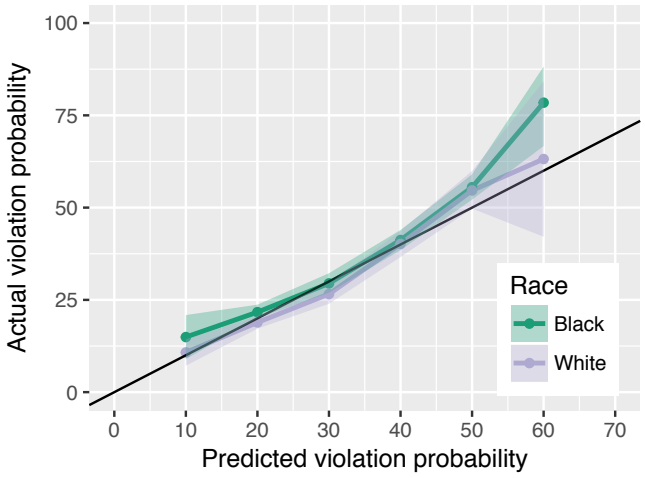

Figure A.1: Comparison of risk assessment predictions and actual violation probabilities for black and white defendants, indicating that the model is well-calibrated across race. Bands indicate 95% confidence intervals.

We then applied the model (i.e., the risk assessment) to make predictions for every defendant in the test set. The model achieves an area under the curve (AUC) of 0.67, indicating comparable accuracy to COMPAS [37, 45], the Public Safety Assessment [13], and other risk assessments [14]. According to a recent meta-analysis of risk assessments, our model has "Good" predictive validity [15].

We also evaluated the risk assessment for fairness. As Figure A.1 indicates, the model is well-calibrated: at every risk score from 10% to 60% (the full range of risks predicted), black and white defendants are statistically equally likely to violate pretrial release. We focused on calibration not as an ideal metric for fairness (recognizing that no perfect metric for fairness can exist [32]), but because it is the most commonly-used approach for evaluating risk assessments in practice [16, 25, 44]. In fact, similarly to COMPAS [3], we find that our model disproportionately makes false positive errors for black defendants compared to white defendants (7.0% versus 4.6%, assuming a naïve threshold of 50%).

Given these evaluations for accuracy and fairness, our risk assessment resembles those used within U.S. courts.

We then selected from the test set a sample of 500 defendants whose profiles would be presented to participants during the experiments in Stage 2 of the study. To protect the privacy of defendants, we restricted our sampling to include only defendants whose attributes along the features shown to participants (the five features included in the risk assessment along with race and gender) were shared with at least two other defendants in the full dataset. While this made it impossible to obtain a perfectly representative sample from the test set, we found in practice that sampling with relative selection weights equal to each defendant's risk score yielded a subset that closely resembles the full set of released defendants across most dimensions (Table A.1).

	All Released N=47,141	**All (Black)** N=26,246	**All (White)** N=20,895	**Experiment Sample** N=500	**Sample (Black)** N=303	**Sample (White)** N=197
Background						
Male	76.7%	77.7%	75.4%	79.6%	81.9%	76.1%
Black	55.7%	100.0%	0.0%	60.6%	100.0%	0.0%
Mean age	30.8	30.1	31.8	28.7	27.7	30.3
Drug crime	36.9%	39.2%	34.0%	42.6%	44.9%	39.1%
Property crime	32.7%	30.7%	35.3%	34.6%	33.3%	36.5%
Violent crime	20.4%	20.9%	19.8%	17.8%	17.8%	17.8%
Public order crime	10.0%	9.3%	10.8%	5.0%	4.0%	6.6%
Prior arrest(s)	63.4%	68.4%	57.0%	54.4%	61.7%	43.1%
# of prior arrests	3.8	4.3	3.1	3.5	4.2	2.4
Prior conviction(s)	46.5%	51.2%	40.7%	35.4%	40.9%	26.9%
# of prior convictions	1.9	2.2	1.6	2.0	2.3	1.5
Prior failure to appear	25.1%	28.8%	20.4%	27.6%	33.0%	19.3%
Outcomes						
Rearrest	15.0%	16.9%	12.6%	15.4%	17.8%	11.7%
Failure to appear	20.3%	22.6%	17.5%	19.6%	19.5%	19.8%
Violation	29.8%	33.1%	25.6%	29.8%	31.4%	27.4%

Table A.1: Demographics and criminal backgrounds for all of the defendants who were released before trial and for the 500-defendant sample used in the Mechanical Turk experiments, broken down by defendant race. A violation means that the defendant was rearrested before trial, failed to appear for trial, or both.

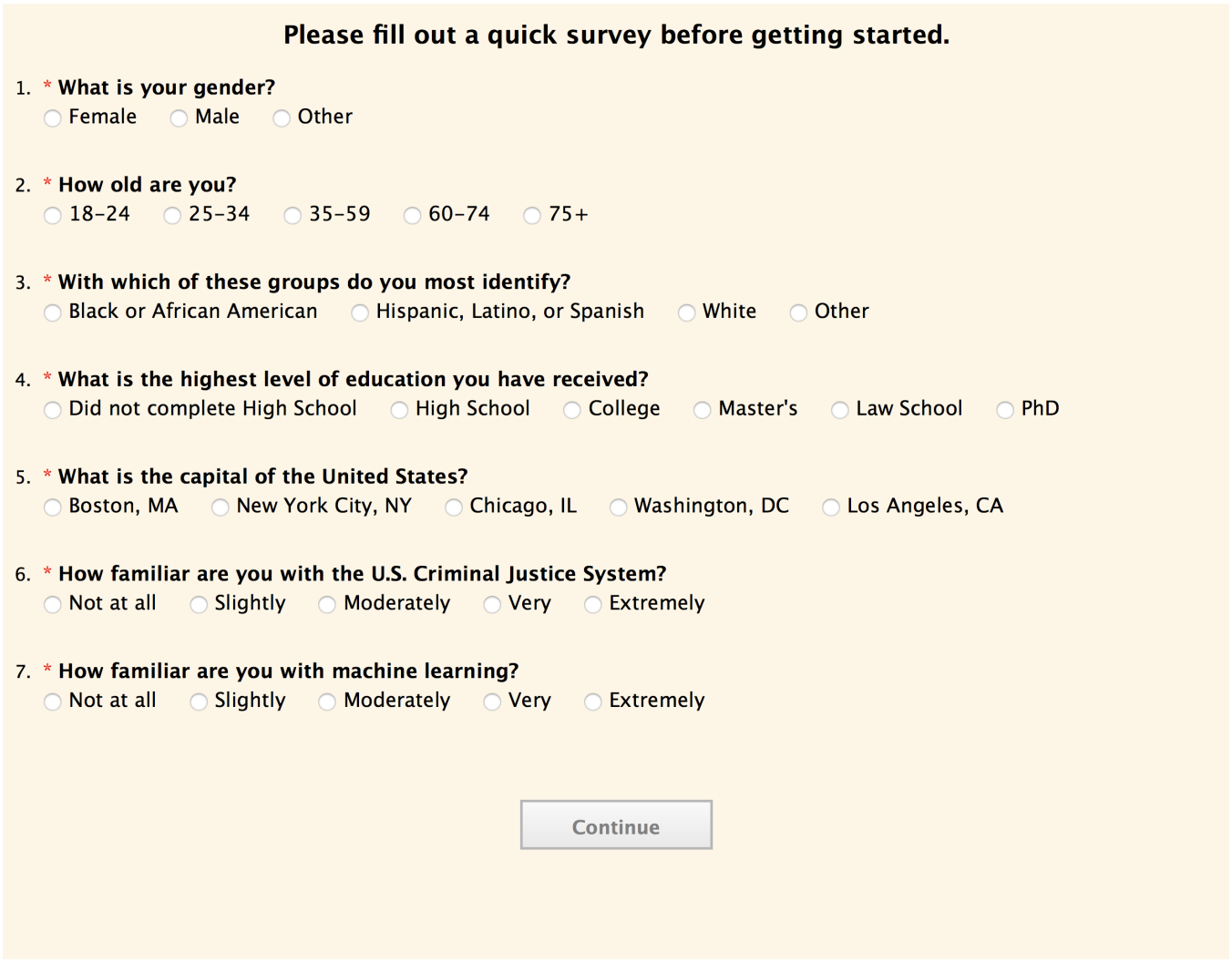

Figure A.2: The intro survey presented to all participants.

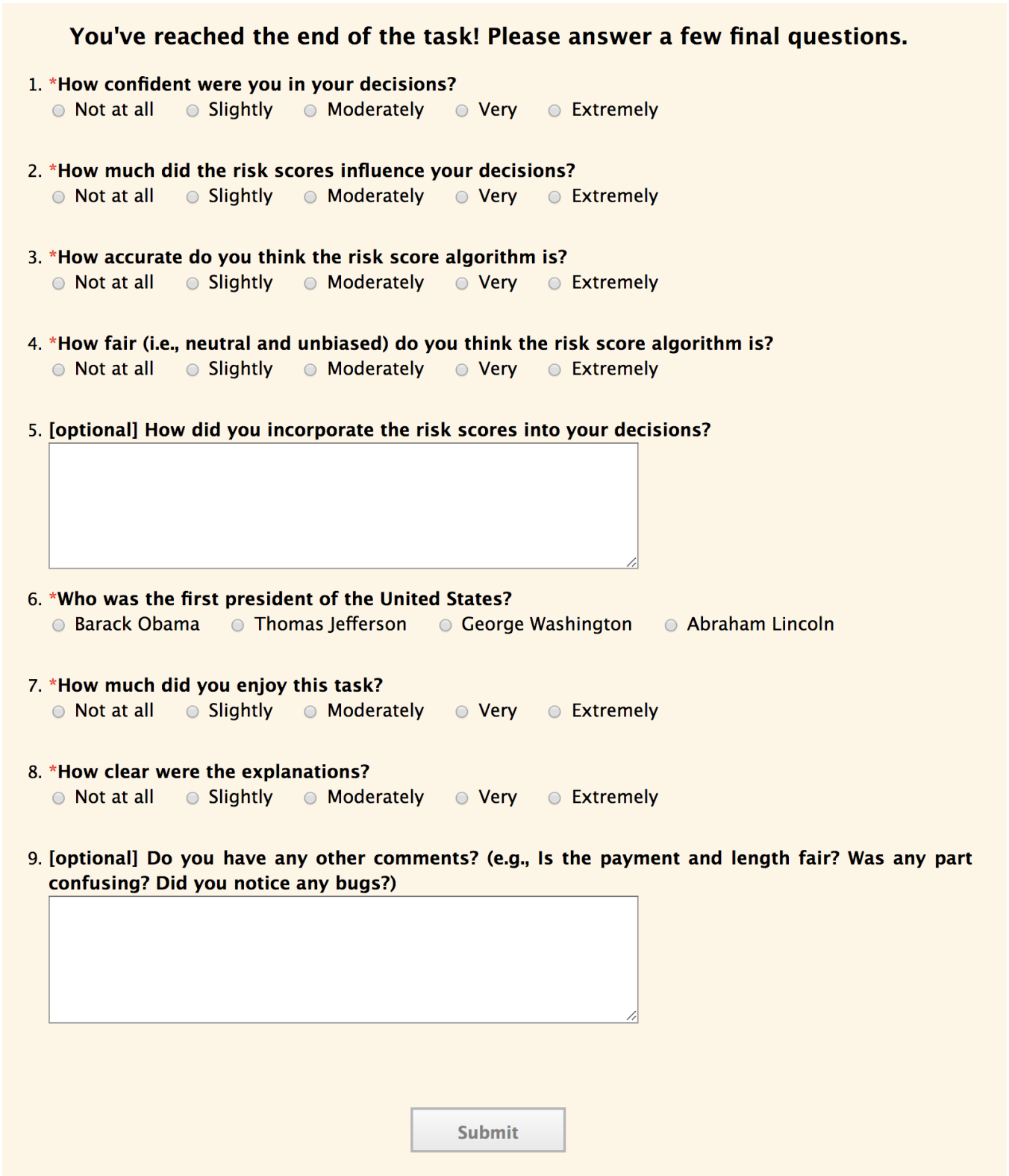

Figure A.3: The exit survey presented to participants in the treatment group. Participants in the control group were not presented with questions 2–5.

	All N=554	Control N=250	Treatment N=304
Male	58.5%	60.4%	56.9%
Black	7.4%	8.0%	6.9%
White	80.5%	80.0%	80.9%
18-24 years old	9.7%	7.6%	11.5%
25-34 years old	42.4%	43.6%	41.4%
35-59 years old	43.9%	44.4%	43.4%
60-74 years old	4.0%	4.4%	3.6%
College degree or higher	65.5%	67.6%	63.8%
Criminal justice familiarity	2.8	2.9	2.8
Machine learning familiarity	2.4	2.3	2.4
Experiment clarity	4.4	4.5	4.4
Experiment enjoyment	3.6	3.6	3.7

Table A.2: Attributes of the participants in our experiments.

Deviated from the risk assessment (N=79, 50.6%; average reward=0.79)
"I used the risk scores as a starting point and then I made adjustments based on my own intuition about each case."
"I used them as an anchor point, and then shifted up or down one depending on my personal feelings about the individual cases."
Incorporated the risk assessment after making own judgment (N=58, 37.2%; average reward=0.79)
"I did not consider it until after making my own decision and then adjusted accordingly."
"decided on a score myself first, then I let the risk score slightly sway my decision."
Followed the risk assessment completely (N=10, 6.4%; average reward=0.81)
"I input exactly what the risk score indicated. It's probably smarter than I am."
"I used the risk score all the time for the entire HIT. Machine learning is much more accurate than humans."
Ignored the risk assessment entirely (N=9, 5.8%; average reward=0.77)
"I just went with my own thoughts after reading each scenario."
"I didn't really pay that much attention to it since I felt the percentages were too low."

Table A.3: A representative sample of the responses that treatment group participants submitted when asked on the exit survey about how they incorporated the risk scores into their decisions (Question 5 in Figure A.3), broken down by the general strategy they indicate having used.

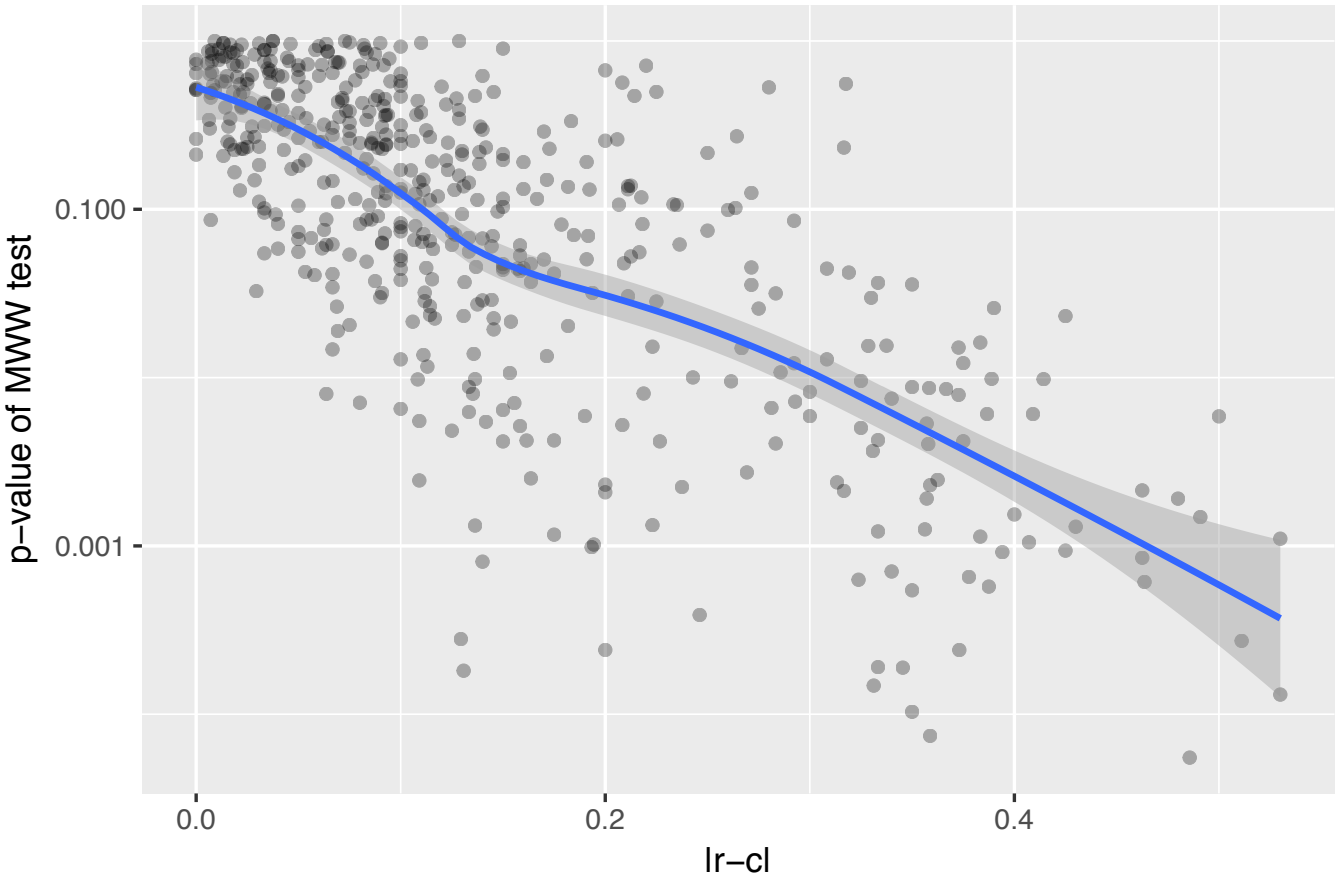

Figure A.4: Comparison of x) the difference between the risk score (r) and the control group's average prediction (c) and y) the difference between the distributions of predictions made by the control and treatment groups, as measured by the p-value of a Mann-Whitney-Wilcoxon (MWW) test. Each dot represents one defendant and is made partially transparent such that darker regions represent clusters of data. The blue line and gray band represent a local regression (LOESS) smoothing fit and 95CI. As r and c diverge, the treatment and control group prediction distributions also diverge. This indicates that, although our analyses focused on the average predictions made by the control and treatment groups, the risk assessment influenced the full distribution of predictions made by the treatment group.

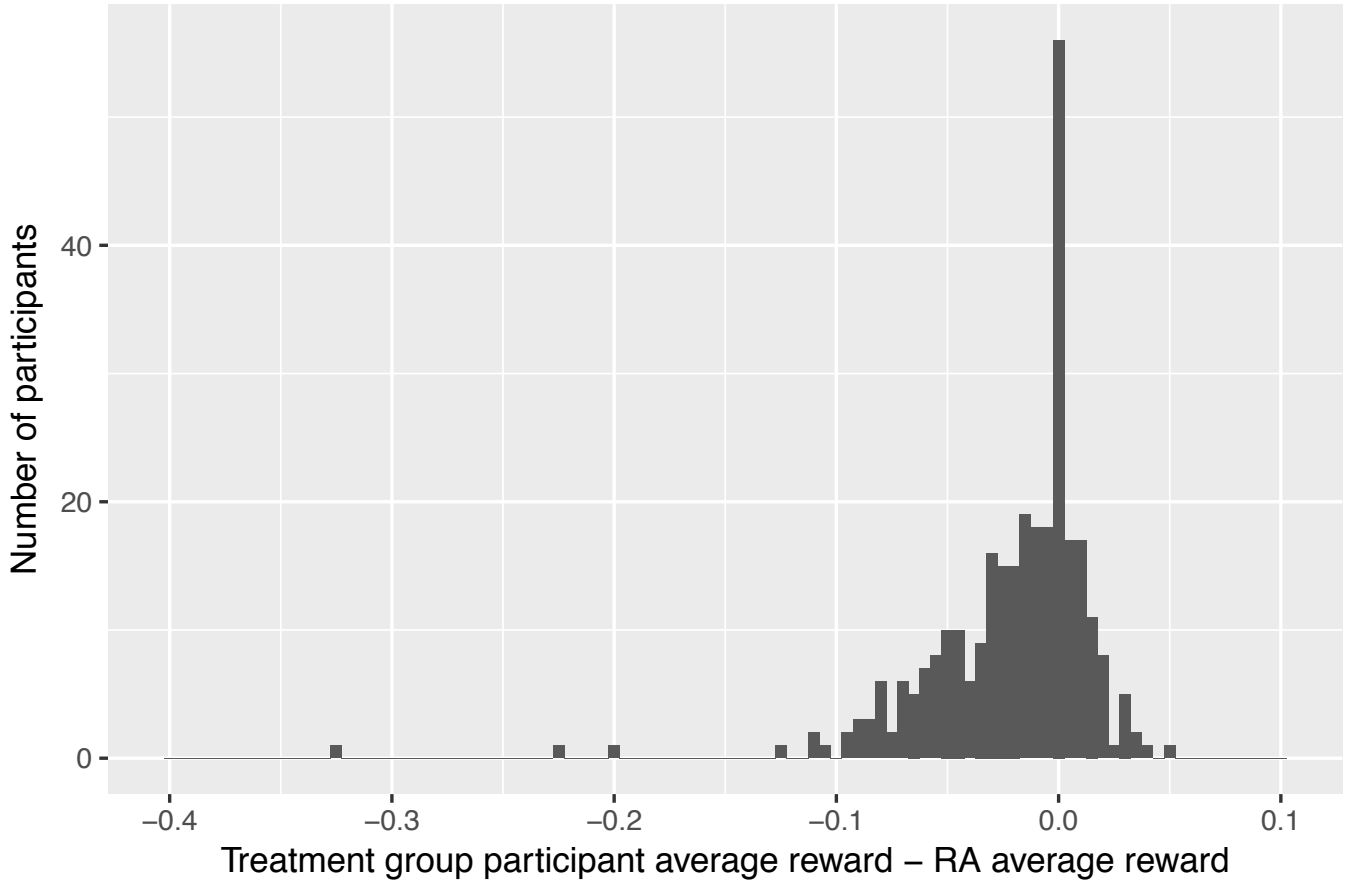

Figure A.5: Distribution of differences between participant performance and risk assessment (RA) performance over the course of each treatment group participant's trial. Negative values indicate that the treatment group participant received a lower average reward than the risk assessment for the 25 predictions that the participant made. Out of the 304 treatment group participants, 195 (64.1%) earned a lower average reward than the risk assessment, 37 (12.2%) earned an equal average reward, and 72 (23.7%) earned a larger average reward.

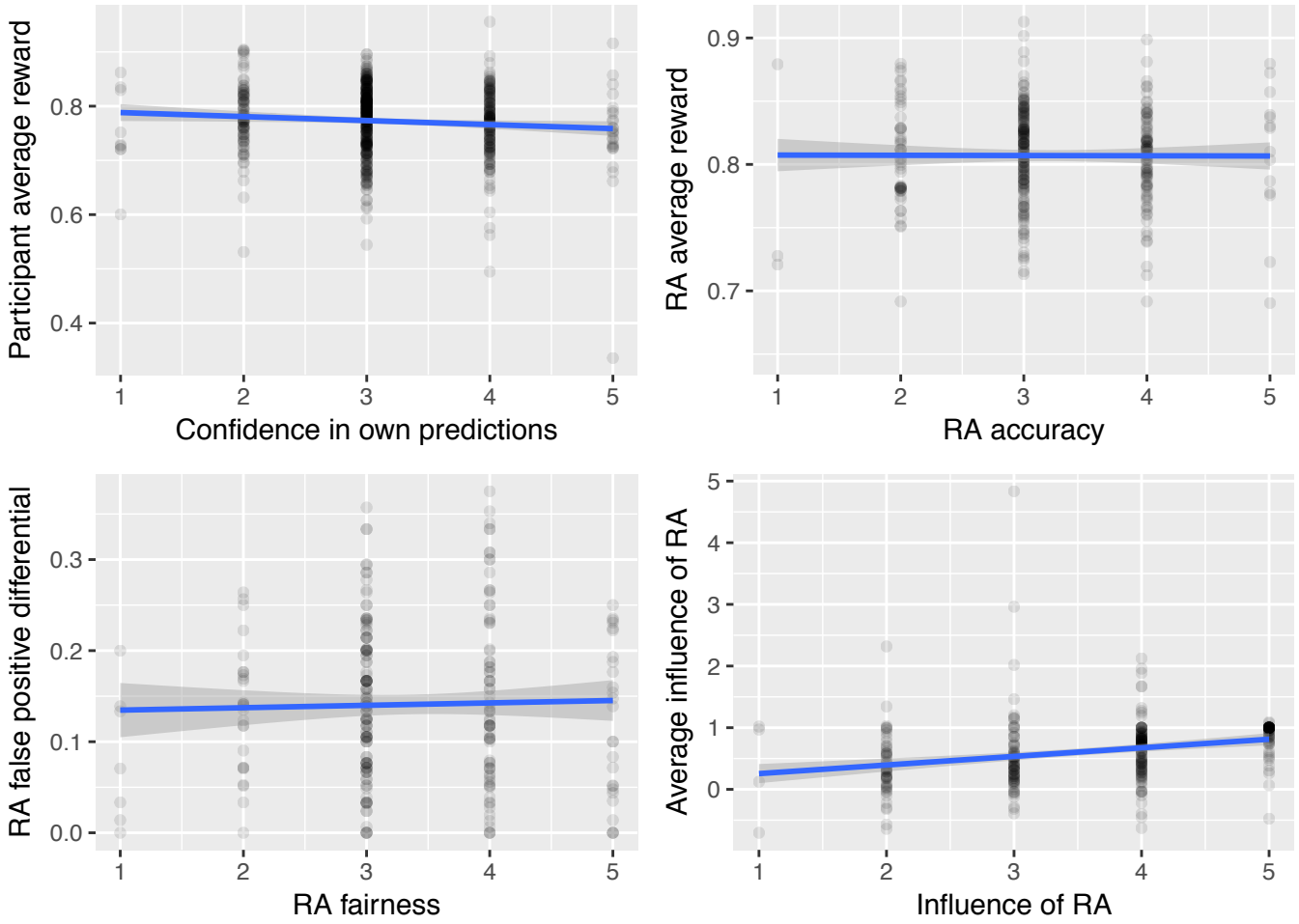

Figure A.6: Comparison of participant evaluations and the actual behaviors of themselves and the risk assessment (RA). Each x-axis represents the participant reflection provided for the first four questions of the exit survey (Figure A.3); the y-axes represent a proxy for the actual outcome that the participant was evaluating (as described in Section 4.3). Each dot represents one participant and is made partially transparent such that darker regions represent clusters of data. The linear regression fits presented here do not include the controls described in Section 4.3, but are shown for demonstration purposes, as the fits depicted closely resemble the relationships found in the full regression analyses.

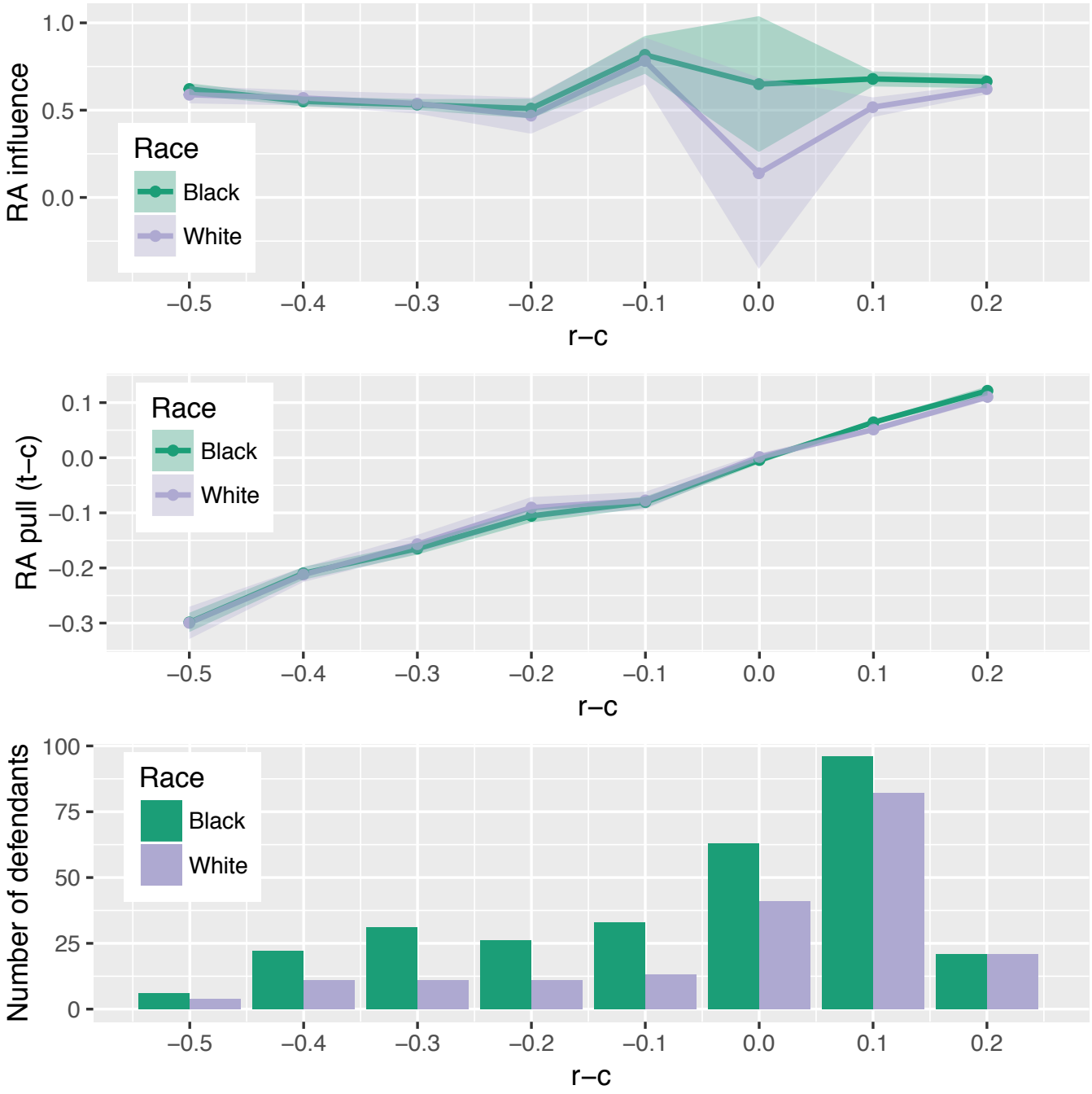

Figure A.7: Top: The average influence of the risk assessment on treatment group participants (as measured by Equation 1) based on defendant race and the difference between the risk assessment and control group predictions (r − c), rounded to the nearest 0.1. The bands depict the standard error for each group; the standard errors around the r − c = 0 groups are particularly large because (given that r − c is the denominator of Equation 1) the influence measurements become unstable when r and c are almost identical (for this reason we excluded the eight defendants for whom r = c from all three panels). The differences in the risk assessment's influence across race are statistically significant only when r − c = 0.1: the average influence on participants evaluating black defendants is 0.68 while the average influence on participants evaluating white defendants is 0.52 (p = 0.02, 95CI difference in means [0.02,0.30]). Middle: The actual change in risk prediction instigated by the risk assessment (i.e., t − c, the numerator of Equation 1). The differences in the risk assessment's pull across race are statistically significant only when r − c = 0.1: the average increase for black defendants is 0.064 while the average increase for white defendants is 0.051 (p = 0.042, 95CI difference in means [0.0005, 0.0255]). Bottom: The number of black and white defendants who fall into each category.