Quantifying Effects of Automated Noise Auditing Notices and Decision Structuring on Noise, Accuracy, and Fairness in Human Decision-Making

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37371753

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Quantifying Effects of Automated Noise Auditing Notices and Decision Structuring on Noise, Accuracy, and Fairness in Human Decision-Making

A thesis presented by
Sahana Srinivasan
to
The Department of Computer Science

in partial fulfillment of the requirements for the degree of
Bachelor of Arts
in the subject of
Computer Science

Harvard University
Cambridge, Massachusetts
May 2022
Quantifying Effects of Automated Noise Auditing Notices and Decision Structuring on Noise, Accuracy, and Fairness in Human Decision-Making

Abstract

Subjective decision-making by a single human decision-maker is pervasive in modern society; the effects of these decisions in domains like criminal justice are extremely significant. Prior work has aimed to improve human decision-making by developing predictive models that assist judges, but such algorithms often imperil fairness or are met with public hesitation. We thus aim to develop alternative ways of improving human decision-making with the assistance of automation and software but not complex algorithms. Specifically, we aim to reduce noise levels in human decision-making to make decisions more consistent, accurate, and fair. We define noise as the random inconsistencies and errors in decision-making that humans are prone to and that deterministic machines, by definition, avoid. In this work, we tested the effects of automated interpretations of various noise-reducing strategies proposed in literature on the noise, accuracy, and fairness of human-made decisions.

We had 300 total participants on Mechanical Turk evaluate criminal defendants’ profiles and predict their risk of re-offending. In our control survey, we gave participants no additional information before they decided. In our first experiment we informed participants of typical noise levels in previous participants’ decisions before they submitted their own predictions, in the second we asked reflective questions to structure the decision-making process before participants submitted decisions, and in our final experiment we calculated participants’ own noise levels using a calibration test and informed them of this noise midway through the experiment. We found that none of our strategies lowered, or impacted at all, noise levels or accuracy in decisions. However, the calibration test actually lowered participants’ positive predictive value from 58.9% to 49.5%. Structuring the decision-making process did improve fairness by lowering discrepancies between participants’ positive predictive value on Black and white defendants. However, providing participants with the individualized, calibrated noise notice decreased fairness by creating a statistically significant difference between participants’ prediction accuracy on white versus Black defendants (52.6% versus 60.2%). Though only the structuring strategy improved group fairness, it was participants who saw the calibrated noise notices who perceived themselves as fairer. This result implies that the calibrated noise notice intervention could dangerously lead decision-makers to falsely believe their decision-making has improved in fairness.

We were thus unable to yet identify a strategy that lowers noise levels, and only one of our strategies improved fairness and none improved accuracy. However, we believe that continuing to test the effects of more decision hygiene strategies on noise, accuracy, and fairness has valuable potential to improve human decision-making, since such strategies can serve as an alternative to machine learning interventions and their drawbacks.
## Contents

0 **Introduction**  
0.1 Our Contributions .................................................. 7

1 **Background and Related Work**  
1.1 Strategies to Minimize Noise ....................................... 9
1.2 Relevant Decision Science Literature ............................ 12
1.3 Decision-Making HCI Literature .................................. 13
1.4 Fairness and Prior Algorithmic Reforms ......................... 15

2 **Methodology**  
2.1 Overview of Methodology ........................................... 17
2.2 Survey Design .......................................................... 19
2.3 Case Data ............................................................... 22
2.4 Participant Recruitment ............................................. 23
2.5 Experiment 1: Noise Notice ......................................... 24
2.6 Experiment 2: Structuring Human Decision-Making .......... 25
2.7 Experiment 3: Noise Calibration Test ............................ 27

3 **Results**  
3.1 Data Overview .......................................................... 29
3.2 Data Quality ............................................................ 33
3.3 Noise Analysis .......................................................... 35
3.4 Accuracy ................................................................. 42
3.5 Fairness ................................................................. 46
3.6 Self-Perception ......................................................... 51

4 **Discussion**  
4.1 Noise Analysis .......................................................... 56
4.2 Accuracy ................................................................. 59
4.3 Fairness ................................................................. 60
4.4 Self-Perception ........................................................ 62
4.5 Future Work ............................................................... 63

5 Conclusion ................................................................. 66

Appendix A Survey Excerpts .............................................. 70

References ....................................................................... 86
Listing of figures

3.1 Group noise metrics ................................................................. 40
3.2 Internal, individual noise metrics .............................................. 41
3.3 Accuracy values ................................................................. 45
3.4 Group fairness results .......................................................... 49
3.5 Confidence and perception of accuracy, fairness ....................... 53

A.1 Survey consent form (constant across experiments) .................. 71
A.2 Opening demographics questions, part 1 ............................... 72
A.3 Opening demographics questions, part 2 ............................... 73
A.4 Task instructions, control and noise notice ............................ 74
A.5 Task instructions, structured and calibration test ...................... 75
A.6 Data quality check questions .................................................. 76
A.7 Example case (control and calibration test experiments) .......... 77
A.8 Example case and noise notice (noise notice experiment) ........ 78
A.9 Example case (structured experiment) .................................... 79
A.10 Example mid-survey, calibrated noise notice (calibration test experiment) 80
A.11 Participant self-evaluation questions, part 1 .......................... 81
A.12 Participant self-evaluation questions, part 2 .......................... 82
For Mom, Dad, and Hansa.
Acknowledgments

First and foremost, I’d like to thank my advisor, Professor Jim Waldo, for his invaluable guidance from the conception of this thesis topic to its final thoughts, his unwavering support whenever I hit roadblocks, and his formative advice on this project and research in general. Of course, I’m also so grateful to my parents, brother, friends, and family for all their support during the last four years and always.
Introduction

People are naturally prone to inconsistencies, variation, and random error in their decisions. Different decision-makers presented with the same situation — even situations with regulated sets of information, such as the profile of a loan applicant or the criminal history of a criminal defendant — often make different judgements. Even the same decision-maker can experience inconsistencies in how they make judgments based on extraneous factors they are unaware of. After all, unlike computers, humans are subject, reasonably, to fatigue from repetition, general fatigue, and emotions. Having to make repeatedly
similar decisions also tend to make decision-makers revert to status quo or mean judgments until their next break in work. A famous example of this phenomenon is a 2011 study that showed that the rate of positive judicial decisions hovers at around 65% immediately after a judge takes a meal break and drops to close to 0% by the end of the subsequent continuous session of work, before the next meal break. This work suggests that factors completely extraneous to the situation at hand can influence decision-makers’ judgments, even unconsciously. Throughout this thesis, we will refer to these seemingly random inconsistencies, variations, and errors in judgement as noise, as shorthand for noise in subjective human decision-making. Our aim in this work is to implement automated, scalable strategies to reduce the amount of noise in subjective human decision-making processes and then test whether they actually do so.

Subjective decision-making is pervasive and typically has high-stakes effects on individual lives and society at large. For example, loan and mortgage applicants are subject to the decisions of individual loan officers (now commonly assisted by credit scores and predictive algorithms). Historically, the extent of this subjectivity allowed loan-granters and banks to participate in the practice of redlining, in which racial minorities — Black mortgage applicants in particular — were regularly denied based on their neighborhood or zip code. As these geographic regions were typically segregated and thus strong proxies for race or ethnicity, this practice constituted race-based discrimination. Redlining therefore denied many Black Americans the opportunity to have a mortgage and pay it off, which is a classic means of building up a credit history and a high enough credit score to secure future financial transactions. So even decisions made decades ago still reverberate and exacerbate inequalities today.

Criminal justice is perhaps the domain in which human judgment has been researched the most by computer scientists in particular in recent years. Major, single-judge decisions are made at several points in the criminal justice system, most notably at the bail and sentencing stages. Criteria for deciding bail differ by jurisdiction, but in Massachusetts, when deciding to let someone out on bail as they await trial, judges are supposed to assess the risk of the defendant fleeing or committing another crime and to not
grant bail release to those considered high risk. Sentencing is the more formal legal consequence assigned to someone who has already been convicted. Academic work and government officials alike have found that judges display implicit biases in these decisions, with some of the latter consequently advocating for reforms including increased implicit bias training. The United States’s criminal justice system has also more broadly faced a reckoning in recent years, with numerous protests against systemic racism and calls for reform or the abolishment of police.

Recent lawsuits against Harvard University and the University of North Carolina that challenge the constitutionality of race-conscious college admissions processes further underscore the tension, anxiety, and importance that surround decision-making processes that significantly affect individual people’s lives. The magnitude of the consequences of these subjective processes on both individual people and the functioning of society means that critically evaluating them and ensuring their effectiveness and fairness is imperative.

Different solutions have developed in different fields to bolster these subjective, human decision-making processes. Risk assessments have become increasingly popular at the local level in the criminal justice system; such assessments are data-driven and are commonly used to standardize the process by which an accused person is evaluated for bail or sentenced after conviction. Some risk assessments have also been developed to decide which activities or programs a person who is incarcerated should engage with or to decide treatment options, restrictions, and supervision policies for people released from incarceration. In a vastly different part of the criminal justice system, algorithmic risk assessments have also been used by family and child protective services organizations to quantify the risk a child faces in their current living situation, in order to decide whether to intervene, remove a child, and potentially prosecute a guardian.

Risk assessments used in bail decisions typically analyze a defendant’s criminal history and general background (such as education level, marriage status, and age) and develop a risk score meant to quantify the predicted risk of that defendant committing another crime or fleeing while on bail. Having a
lower risk score as assigned by the risk assessment thus indicates that letting a defendant out on bail is a more viable option. Typically risk assessments do not actually make any final decisions; a person assigned a high risk score by an algorithm is not automatically denied bail. Instead, human judges use the output of these risk assessments as yet another input to their final, binding decision.\textsuperscript{15} Though risk assessments have existed in some form since the 1920s, only recently have they become especially pervasive and started taking the form of algorithms, which are typically advanced machine learning models.\textsuperscript{14}

For example, the infamous COMPAS risk assessment software created by Northpointe, Inc. analyzes defendants’ responses to a thorough, 137-prompt questionnaire to predict their risk of recidivism — committing another crime — and their risk of committing violent future crime.\textsuperscript{17} The COMPAS questionnaire includes not only questions on a defendant’s current charges and criminal history but also a self-evaluation of their personality traits and how they handle emotions, and questions on their family’s criminal history, their hobbies, habits at school (such as how often the defendant had conflicts with teachers), and more.\textsuperscript{17}

Some people view risk assessments like these as reform, particularly since the cash bail process by which a dollar amount of bail is set and must be paid for the accused to be release pending trial is increasingly being viewed as discriminatory and thus dismantled.\textsuperscript{18}\textsuperscript{19} Over 60\% of the United States population already currently lives in a jurisdiction that uses some sort of risk assessment tool in their criminal justice system.\textsuperscript{20} Again, many of these are now algorithms or advanced machine learning models.

These assessments are representative of a broader modern phenomenon in which algorithms are brought into a process to make it some combination of easier, quicker, cheaper, or otherwise ostensibly better than it would be if it was otherwise conducted either partially or primarily by a human.\textsuperscript{21} After all, algorithms will not fatigue after too many cases and revert to mean judgments, be unduly influenced by unknowable factors in their lives extraneous to the cases they are deciding on, or introduce random error to their decisions unless the algorithm itself intentionally involves randomness.\textsuperscript{1} Algorithms often do out-perform human decision-makers with respect to raw accuracy.\textsuperscript{22}
However, as more and more algorithms like COMPAS have been deployed in critical areas of society --- not just in the bail setting or the criminal justice system as a whole but also in deciding welfare benefits and curating social media feeds, for example --- concerns about the full impact of such tools on society and on the real people they evaluate have also proliferated, particularly since these tools are ostensibly built only to improve processes in the domains in which they are used.²¹ Most infamously, ProPublica investigated the COMPAS algorithm and published an article asserting it discriminated against Black Americans,²³ though the creators of the algorithm and other independent researchers have asserted, conversely, that it does not and that ProPublica’s results were mathematically inevitable due to a fairness impossibility theorem.²⁴ Other, more agreed-upon instances of unfairness or bias in algorithms have been found in deployed algorithms: researchers in 2019 found that a popular risk algorithm meant to predict which patients need the most medical care was biased against Black patients, for example.²⁶ Additionally, Amazon used to use a hiring algorithm that predicted which job applicants would become the best-performing employees. This algorithm was later found to discriminate against women.²⁷

The causes of these algorithmic biases are multi-fold: it could be a lack of representativeness in the training data of a machine learning algorithm, or that the creators of an algorithm have imbued their own biases intentionally or unconsciously into their work. An algorithm could intentionally discriminate based on protected attributes like race, or the cause of bias could simply be a lack of attention to how an algorithm interacts with other parts of its domain, meaning that the algorithm itself may not be biased but its deployment methodology makes it so.²⁸²⁹ Some of these issues are general, such as a lack of attention to how components of a decision-making system interact, and would persist even if we removed algorithms from decision-making pipelines and focused on improving human decision-making in some other way. Some of these issues would simply be replaced by others, such as decision-makers themselves potentially having significant biases. However, the concerns around fairness in predictive algorithms are real and widespread enough that we think it valuable to look into alternative ways to improve human decision-making, which is what this thesis begins to explore: ways to use software — but not
intelligent algorithms — to automate strategies that reduce noise in subjective human decision-making.

Additionally, the discovery of bias and unfairness in technology even completely unrelated to predictive algorithms still affects the viability of these tools as effective solutions to improving human decision-making, since such news erodes already-low public trust in black-box technologies like these complex machine learning algorithms. For example, in 2018 Buolamwini and Gebru found that state-of-the-art, publicly released facial recognition software developed by major tech companies performed significantly worse at identifying people with dark skin than people with light skin. Though people are not inherently suspicious of all algorithms, they do often lose trust immediately after they learn of any mistakes.¹

According to public sentiment studies done by Pew Research, people are now broadly concerned that algorithms deployed in the real world may violate privacy, that they could be unfair, and that they could rob situations of the healthy nuance and insight provided by human decision-makers.³¹ 56 percent of those surveyed said using algorithms in criminal risk assessments to evaluate people up for parole was unacceptable.³¹ To align with the public desire to have some human-provided nuance in decisions, we do not aim to eliminate disagreement, simply to minimize random errors.

Thus, subjective decision-making is everywhere, and it often has serious, significant impact on people’s lives. The main weakness of this kind of human decision-making that we focus on in this thesis is that it often results in a lot of noise — random errors, variations, and inconsistencies — since people are prone to decision fatigue and unconscious influence by factors extraneous to the situation at hand.¹ Algorithms like risk assessment software have been deployed increasingly often in such subjective decision-making contexts, particularly in the criminal justice field, to support the human decision-making process and alleviate the impact of weaknesses including bias and noise, since algorithms fully eliminate noise. But a little noise can be healthy, and algorithms can additionally have their own issues with bias and a lack of transparency; the public is also often mistrustful of such tools, especially those that are hard to explain fully, lack transparency to the public, or that have previously made critical mistakes. Already sixty percent of Americans live in jurisdictions subject to criminal justice risk assessment algorithms.²¹
Thus, since the general public is subject to the consequences of these algorithms, which can be as sig-
ificant to their lives as how long they spend in prison or where they attend school, public comfort and
approval should play a role in what decision making processes look like. People tend to be more trustful
of and comfortable with human decision-making — even if they are aware that humans make mistakes
as often or more often than machines.¹ For this reason, and due to the widespread nature of important,
subjective human decision-making, and because of the aforementioned flaws and perils of current algo-
rimthic solutions that aim to assist human decision-making, pursuing methods of improving of human
decision-making without involving machine learning algorithms is a valuable research path.

0.1 Our Contributions

In this thesis, we will do exactly that: seek to improve subjective human decision-making processes with
the assistance of some software and automation but without the use of complex algorithms or artificial
intelligence to actually make the decision. Improvement is a general goal that could be implemented in
many ways — making decisions more efficient or ensuring greater fairness, for example. In this work,
the axis of improvement that we will focus on is noise. We will specifically test whether formalized, semi-
automated versions of the noise-reducing techniques proposed by Kahneman et al. in their book, Noise,
actually do reduce noise in decisions without compromising — or perhaps even while also improving —
accuracy and fairness.

We believe reducing noise is a desirable goal, one that will improve subjective human decision-making,
for several reasons. Firstly, a lower level of noise means that decisions will likely be more interpretable
and understandable, since there is less random error clouding them. Additionally, lower noise levels
likely indicate that factors extraneous to the situation being evaluated are having a lesser impact. Ad-
ditionally, although the concept of fairness in computer science literature is a separate and complex
concept, reducing noise can itself ensure greater fairness in a system by reducing random variation in
judgment and thus ensuring that similar people are treated similarly. Treating similar people similarly is one of the most popular definitions of fairness in computer science.  

If we do validate such a method of reducing noise in human decision-making without introducing the problems associated with predictive machine learning algorithms, then we could ideally — in real time as real decision-makers pass judgments — help them act more consistently, more fairly, and with less random error, improving the performance of the system and people’s treatment by the decision-making systems.

Chapter 1 of this thesis describes relevant background and related prior work, including more detail on what noise is and why lowering it is beneficial, what Kahneman et al.’s proposals for lowering it are, and prior efforts to improve subjective human decision-making. Our contributions are contained in chapters 2-5: chapter 2 outlines the methodology of our experiments and our hypotheses, chapter 3 summarizes our results, chapter 4 discusses the implications of these results for our original goal and hypotheses, and chapter 5 offers some concluding thoughts on this line of research.
1.1 Strategies to Minimize Noise

The methodologies used in this thesis to identify an effective way of reducing noise levels in human decision-making are inspired by the 2021 book *Noise: A Flaw in Human Judgment* by professors Daniel Kahneman, Olivier Sibony, and Cass Sunstein. They present the definition of noise that we use in this thesis as inexplicable variations — random errors, caused by humans not making decisions with precise
mathematical rules as algorithms do, which means that mood or time of day or other unidentifiable, extraneous factors can influence judgment. In practical decision-making, noise becomes apparent when one same case evaluated by several decision-makers receives significantly different judgments or when a single decision-maker makes significantly different judgments on similar cases.¹

The authors of the *Noise* book make an immediate value judgment of noise as something undesirable and to be mitigated. However, we acknowledge that noise is not universally seen as a negative in decision-making. Some argue that focusing too much on removing noise could result in throwing out useful, nuanced information whose true value or underlying patterns is simply not yet understood, causing researchers to lose out on future insights. Deciding what noise even is, is subjective and could itself be a biased decision-making process.³³

Nevertheless, we make the same assumption as Kahneman et al. of noise as something that we aim to reduce in judgement because of the close ties between lower levels of noise and greater individual fairness, which is a concept that computer science as a broader field has accepted as desirable and which will be introduced in full later. In short, individual fairness is defined as a system treating similar people similarly; the opposite of this phenomenon, similar people being given different or a wide variety of judgments, is seen by Kahneman et al. as indicative of the presence of noise.³² Consistency — which is the opposite of having high levels of noise in a system — is also valued in decision science literature; it is as central to the field’s work as accuracy is.³⁴

Deterministic algorithms have no noise: if you re-run the same machine learning model on the same input at a different time or on a different day, the algorithm will still output the same result.¹ Machines being noiseless in this way is part of the motivation underlying our work: since one important benefit of machines as decision-makers is their lack of noise, and their fairness instead is what concerns many, then maybe machines — or some form of software or automation — can be used to reduce noise in human judgment, without touching the rest of the decision-making process or making any decisions outright, thus avoiding the fairness and bias concerns discussed earlier.
The core of Kahneman et al.’s work that we are building off of are suggestions on how to replicate the noiseless aspects of machines in humans. Some prior research has already tried to do this: Goldberg in 1970 found that linear models built to predict what human judges would decide in a case, rather than predict the actual future outcome of a case, could actually predict outcomes more accurately than the judge they were originally emulating. Kahneman et al. hypothesize that this happened because a linear model captures the gist of the judge’s logic but removes subtleties, and some of those subtleties were undoubtedly causing noise without improving judgement. Additionally, models lack any of the emotions or extraneous factors irrelevant to the case that could be unconsciously influencing a human decision-maker. This research by Goldberg, however, is still an example of an algorithm entirely replacing humans in the decision-making process, which, due to questions around machine fairness and transparency and skepticism from the public, we wish to avoid in our work. We additionally wish to avoid situations in which an algorithm makes a decision and presents it to a human decision-maker to make the final call, for prior work has found that in situations like these, human decision-makers tend to anchor their judgment to what the machine told them.

Kahneman et al. propose more than a dozen strategies to reduce noise in human decisions. This thesis will focus on only some of his solutions — the ones that have the best potential to be automated and thus smoothly incorporated into judges’ real-time decision making processes, the way that risk assessment machine learning models currently are. Again, though we are considering ways to use software to automate these strategies, we do not aim to build any complex decision-making models; such assistive models with their potential bias concerns are what we are trying to find alternatives to.

As a first step to reforming noise levels in any organization, Kahneman et al. recommend a noise audit to help people understand the power and prevalence of noise. Specifically, in this noise audit, an overseer would fix a set of cases specific to the organization’s domain (e.g. a set of loan applicant profiles if this is a loan-granting institution), present them to several different decision-makers, and then quantify the variation among their evaluations of the case. Since most of the rest of his solutions rely on decision-
makers engaging in structured, complex discussions to reduce noise, the noise audit is one of the main solutions we seek to adapt and automate. We will do so by calculating and informing decision-makers in various different ways about the noise levels that typically exist when judges evaluate the recidivism risk levels of criminal defendants. Then we will evaluate the effect of these notices on various properties of decision-makers’ subsequent own risk ratings of criminal defendants, including the noise level among these decisions and fairness metrics. The other idea from Kahneman et al. that is incorporated into our methodology is that complex judgments, such as evaluating the risk level of a criminal defendant, should be broken down and structured. To that end, one of our experiments breaks down the decision-making process and asks for insight into participants’ decision-making processes before prompting them to make a decision.

Other, more broad recommendations from Kahneman et al. on how to minimize noise are to not expose judges to irrelevant information early in the decision pipeline, to aggregate multiple judges’ decisions when possible, and to have judges share a well-defined scale (e.g. sentencing guidelines that roughly correspond the nature of a crime to a range of sentence lengths). Since these techniques apply to elsewhere in the decision-making pipeline from when an individual judge makes their decision, we do not test them in the scope of this thesis.

1.2 Relevant Decision Science Literature

Prior work in psychology has found altering human decision-making processes to be difficult. Much of this work has focused on eliminating biases in human judgment. Milkman et al. in 2009 summarized the state of efforts to rid human decision-making of bias: informing people of bias has proved unsuccessful, and even more extensive, personalized training and coaching has shown only minimal effects. One hypothesis for this lack of effect is that people are not able to internalize the implications of such warnings enough to change their behavior accordingly, especially if they are relying on quicker, intuitive forms of
Some more successful strategies to reduce biases include asking people to think through the opposite of their decision before they make it, which has been shown to reduce bias-caused errors in judgment; making the decision-maker accountable in some way for the outcome of their decision; and having groups jointly deliberate a case. The last strategy is one Kahneman et al. also recommend in *Noise*, though it is not one we are testing in this work.

Several prior studies in decision science have shown that incentives, feedback about past performance, and direct teaching — particularly about unintuitive concepts — make for the most successful interventions in decision-making in general. In line with both this work and Kahneman et al.’s, we incorporate incentives into all our experiments, and one of our experiments is centered around the principle of providing participants with feedback about their individual inconsistencies and noise levels.

### 1.3 Decision-Making HCI Literature

Previous research has also sought to evaluate the effect of machine-involved interventions on human decision-making. Vaccaro in 2019 studied the effect of seeing the COMPAS algorithm’s risk assessment scores on judges’ final assessments of defendants’ risk of recidivism, as well as the effect of seeing these COMPAS risk scores accompanied by various notices explaining the score or providing disclaimers about potential biases in the score. This work found that risk scores acted as anchors to human judges’ scores, with lower risk assessment scores leading judges to provide lower scores themselves, though judges were not able to recognize this happening. Further, Vaccaro found that COMPAS was more accurate at predicting recidivism than human judges were. Additionally, judges seeing the COMPAS scores before they submitted their own risk scores did not significantly affect judges’ overall accuracy, though it did impact fairness metrics of their aggregate decisions, favoring white defendants. Judges who viewed highly technical written notices about the nature and proper uses of the COMPAS algo-
algorithm before predicting themselves had lower prediction accuracy than those who simply viewed the
COMPAS score; this was a statistically significant result. However, when judges viewed an explanation
of the COMPAS algorithm written in more transparent lay terms, based on one that the Wisconsin
Supreme Court required to accompany risk assessments, they performed no differently than people who
saw no explanation. 36 Multiple other studies have found that various different methods of explaining
risk assessment algorithms to judges did not significantly improve final prediction accuracy. 38, 39 Vaccaro’s
work to evaluate the effect of algorithms and disclaimers about algorithms on the accuracy and
fairness of decision-making in criminal justice, in conjunction with Kahneman’s strategies on how to
potentially reduce noise in decision-making, serve as the main points of inspiration for this thesis.

These studies question the validity of the argument that human judges act as a check on the power of
algorithmic risk assessments because humans have the final decision, since COMPAS scores skew judges’
scales more than judges even realize and can actually make judges’ decisions more unfair. Even provid-
ing explanations of risk assessment scores does not compensate for this effect on judges. These results
provide further motivation for research into how we can improve human decision-making without in-
volving predictive algorithms.

Green in 2020 found some slightly contradicting results on similar experiments: presenting the re-
sults of a risk assessment to judges improved their accuracy, but their accuracy still remained lower than
that of the assessment. 41 Further, this work found that people placed more weight on the risk assess-
ment’s output than on their own judgment and that the risk assessment was increasing judges’ accuracy
by actually altering the nature of their decision-making process, making them more attentive to risk in
general, not just because it was an additional, useful piece of input information.

Green also proposed three “desiderata” for joint algorithm-human decision-making processes (coined
as “algorithm-in-the-loop” settings): accuracy, reliability, and fairness. Surveyed judges struggled to
maintain reliability and fairness after seeing COMPAS risk scores, finding it particularly difficult to cal-
ibrate their own judgement to the assessment. 41 The noise metrics that we work with in this thesis are
a combination of reliability and fairness, since lower noise levels correspond both to greater consistency in judgment and in similar individuals being evaluated similarly. Thus, Green’s work implies a potential incompatibility between judges using risk assessment results in their decisions, which they will not be in our experiments, and them maintaining low levels of noise, which is the goal of this thesis.

1.4 Fairness and Prior Algorithmic Reforms

A vibrant sub-field of computer science aims to build fair algorithms and head off the concerns outlined earlier about algorithms that actively discriminate or that quietly perpetuate existing biases. Some approaches to doing so in recent literature include trying to ensure that group fairness metrics hold across all groups in a population that can be identified in a computationally reasonable amount of time, instead of just explicitly protected groups, and using tools from causality literature to estimate counterfactuals and draw fairness conclusions from comparisons between real outcomes and these counterfactuals. An example of the latter work is a 2018 study that demonstrated that even just policy changes to how an algorithm is used in the U.S. criminal justice system could reduce crime rates by 24% without increasing jailing or reduce jailing by up to 41% without increasing crime rates.

Our work explores alternatives to the methods of this field by trying to improve decision-making with algorithms removed entirely from the process. However, to benchmark the quality of the decisions that result from our experiments against the decisions that COMPAS produces, that COMPAS-assisted judges produce, and that newer decision-making algorithms in research produce, we will evaluate our experiments using standard computer science literature metrics of group fairness.

There are two main conceptions of fairness currently used in computer science literature: group or statistical fairness and individual fairness. Group or statistical fairness, which we will refer to in this thesis just as group fairness, measures the fairness of decisions by seeing if certain statistical properties hold. One of the most commonly-used of these properties is classification parity, which stipulates that
the metrics like the false positive rate or the false negative rate are equal across protected groups such as race or gender. Precise equations for this and other metrics involved in evaluating group fairness will be articulated in our Results section as we analyze our experimental data. One major caveat to the reasonableness of using group measures of fairness is that unless you have a perfectly accurate predictor, it is mathematically impossible to satisfy certain of them simultaneously if the protected groups at hand, such as different racial or gender groups, have different true rates of being positive in the prediction task.

Individual fairness is the main alternative conception of fairness; it says that individuals who are similar — e.g. criminal defendants who have similar criminals histories, if that is how the criminal justice system chooses to define similarity — should be treated similarly by a decision-making process. The main challenge in ensuring that algorithms adhere to this definition is defining what similarity means, a task which is domain-specific.

As mentioned before, our goal of minimizing noise can be thought of as also advancing the goal of individual fairness since noise, by definition, is random error that can cause similar cases to be treated differently for no explainable reason. In our methodology, we will even propose a definition of what make cases similar so we can later analyze any differences in how similar criminal defendants are evaluated in our experiments.
2.1 Overview of Methodology

Our guiding research goal is to identify techniques that reduce noise in human judgment without compromising the fairness or accuracy of decisions. As a starting point for this question, we are testing the effects of versions of the solutions proposed by Kahneman et al. in *Noise*, such as making human decision-makers aware of noise before they submit decisions or imposing greater structure on their
decision-making process. However, there are some notable differences between the implementation of the interventions originally posed by Kahneman et al. and the methodology we are testing. One of Kahneman et al.’s two main solutions to reducing noise was a one-time, manual noise audit that leaders of an organization would perform to snapshot the current amount of variation in their team’s decisions. He hypothesized that making decision-makers aware of the typical level of noise in their judgements would help people behave more consistently in the future.¹ His other solution was a decision-sanitizing checklist, a manual, involved process in which an observer would monitor group discussion about a decision, looking for best practices. There is not much benefit to directly automating the first, one-time noise audit, and the second solution involves enough high-level conversation to likely make automation an ineffective implementation technique.

Our experiments thus pulled aspects from both of these solutions: we built a set of surveys that partially automated the strategies of (a) calculating and making participants aware of noise in real time, in two different ways, and of (b) imposing structure on decision-making processes via targeted questions about how decisions were made. We want to be clear about why we want automation and software playing a role in our work, the difference between automation and algorithmic intervention, and the difference between what we are testing and fully automated strategies ready for real-world deployment. To the first point, we seek to automate Kahneman et al.’s strategies and not simply test, for example, the manual one-time noise audit because we ultimately hope to find a strategy to reduce noise that we can scale well and use continuously to check judges as they actively make new decisions. Automation lends itself to scaling. We do not seek to build any complex software that makes decisions itself, only to propose and test a solution that automates noise-reducing strategies. Finally, though the reason we desire an automated solution is to have it scale in real time (e.g. judges are notified of inconsistencies in their decisions or of the potential influence of extraneous factors like time of day as soon as they submit an submit to the software and are then allowed to re-submit a new decision, if that were found to be an effective noise-reducing strategy), we would be unable to test such software in the scope of this thesis,
since it would need real-world deployment, large samples of judges, and contextual information on their decisions such as time of day and schedule to meaningfully evaluate its efficacy.

Thus, we are instead taking a first experimental step of sending out surveys that, via embedded Javascript code or notices about pre-calculated noise levels that we write manually into the survey (which software would calculate live in real deployment), test out partially automated versions of Kahneman et al.’s strategies. If our results show positive effects from these strategies, future work would then fully automate and test these strategies so that judges interacted with directly with software that analyzes them live and has contextual information on them, not surveys. Any distinctions between Kahneman et al.’s original strategies, the semi-automated interpretations in our experiments, and the fully automated versions we envision in future work will be illustrated in more detail when we describe our three experiments in sections 2.5-2.7.

2.2 Survey Design

Our IRB-approved study consisted of four surveys — one control and three experimental. In each survey, participants were given information about criminal justice cases and asked to rate criminal defendants’ risk levels and to predict the likelihood that each defendant would commit another crime, or reoffend, within the next two years. This style of questioning was modeled after prior work that sought to evaluate the fairness and accuracy of human decision-making, machine decision-making, or a combination of both. 21 36

Each of the four surveys followed the same rough format:

- A consent form which included information on the purpose of the research, the fact that participants can later ask for their data not to be used, the contact information of researchers, and compensation instructions. At this point, our consent form disclosed only the vague purpose of our research — studying decision-making processes in criminal justice — to avoid giving partic-
ipants enough information about our treatments and hypotheses to skew their decisions. The form then disclosed to participants that the full nature of our research was being withheld and would be revealed in full after the study, which it was.

- A set of orienting demographic questions, including age bracket, gender, race, highest level of educational attainment, and familiarity with the criminal justice system graded on a five-point scale.

- Task instructions, the details of which vary slightly based on the survey. Across all surveys, the instructions stated that participants would be presented with 40 cases about crimes committed in the United States and that they would be asked to rate the risk level (on a scale from 1 to 10, where we tell them which numbers correspond to low, medium, and high risk) of the defendant in each case and predict whether they would commit another crime in the next two years. For each of the three experimental treatment surveys, the instructions included a brief description of the experimental intervention we were going to present participants with alongside the cases — a notice about noise levels, for example. The instructions were followed by a comprehension question, which was constant across all cases and tested participants’ understanding of what they were supposed to be predicting in the survey.

- 40 defendants’ cases, with the following information provided for each: race, sex, age, criminal charge, criminal degree of charge, non-juvenile prior conviction count, juvenile felony count, and juvenile misdemeanor count. The information on each defendant was followed by two questions about the defendant: one to score their risk level, and one to predict whether they will commit another crime within two years. Other information or questions were included alongside the cases for the three experimental surveys, the details of which will be described in Sections 2.5–2.7. The language of the scoring and prediction questions is pulled from the experiments done by Vaccaro. The exact manner in which cases are described to participants is as follows, also using
the same language as Vaccaro: 36

The defendant is a [RACE] [SEX] aged [AGE]. They have been charged with: [CRIME CHARGE]. This crime is classified as a [CRIMINAL DEGREE]. They have been convicted of [NON-JUVENILE PRIOR COUNT] prior crimes. They have [JUVENILE FELONY COUNT] juvenile felony charges and [JUVENILE MISDEMEANOR COUNT] juvenile misdemeanor charges on their record.

For lesser-known criminal charges, such as “uttering a forged instrument,” brief parenthetical descriptions of the charge were given, such as “knowingly selling, publishing, or passing on a forged financial document with the intent to defraud.”

At a random point in this set of 40 case evaluations we displayed an “attention check” question that told participants which option in a dropdown menu to choose. We posed this question to help weed out low-quality responses.

• The cases were followed by a post-questionnaire that asked participants to rate their confidence in their decisions and the fairness and accuracy of their predictions on a seven-point scale. These questions were also modeled after Vaccaro’s surveys36 to better directly compare the effect of our intervention against those in prior work. A free-form question also allowed participants to provide feedback on the study questions. Participants were finally asked their opinion of the criminal justice system. This question was asked last in the survey to ensure prompted reflection on their opinion of the system did not unduly affect participants’ risk assessments and predictions.

Results from the three experimental treatment surveys were compared against the same set of results from our control survey.
2.3 Case Data

We chose to have participants evaluate criminal justice cases in our experiments for several reasons: firstly, to maximize the likelihood of participants being generally familiar with the domain; secondly, since we want to analyze a domain in which machines are already being deployed to assist or replace human judgment and in which the fairness of decisions is critical; and thirdly, since we had access to real-world data on criminal justice cases, including real-world reoffending outcomes, and could therefore analyze participants’ accuracy.

Our cases were sampled from a dataset of crimes committed in Broward County, Florida between 2013 and 2014. This dataset, for each crime, includes names and demographic information for the defendant as well as their juvenile and non-juvenile conviction history. Our study removed names and other explicitly identifying information from the dataset before use. The dataset was cleaned, organized, and published by ProPublica as part of their work analyzing the fairness of the COMPAS recidivism prediction algorithm. This dataset is thus commonly used or referenced in computer science literature.

When sampling our 40 cases — which were held consistent in each of the four surveys — from the full Broward County dataset, we excluded cases for whom illegality of the charge is inconsistent across states or otherwise now obsolete, including charges related to marijuana, as was done in prior work. We also excluded cases for which we had incomplete information. We randomly sampled 30 cases from the dataset and manually selected an additional 10 that were deemed “similar” to one of the 30 randomly sampled cases. We defined two cases as being similar if the defendants differed in only race or gender but not both, faced charges of the same degree, and for each type of prior conviction (non-juvenile, juvenile felony, and juvenile misdemeanor), either both had 0 priors or both had non-zero prior counts that were off by no more than 1 for juvenile priors and 2 for non-juvenile prior counts. This manual selection step was done to ensure that there were enough pairs of similar cases within the dataset for us to quantify the noise of how differently participants judged similar cases.
2.4 Participant Recruitment

Survey participants were recruited via Amazon’s Mechanical Turk, an online task crowdsourcing platform. Participants were presented with a description of our task, among others, and then self-selected into participating. After selecting one of our tasks, participants were linked to our survey on Qualtrics. Upon completion of the survey, participants were provided with a code to enter into the Mechanical Turk task to indicate their completion of the study. Participants were prevented from filling out multiple of our surveys. Only participants registered on Mechanical Turk as adults living in the United States were able to participate in our study. For three of our surveys — the control, noise notice, and calibration test ones, which will be described in the next section — participants were paid $1.50 via Mechanical Turk’s payment system for completing the full survey. For completing the structured experiment survey, participants were paid $3.00. This discrepancy is because the latter survey is more time-intensive, involving short-answer questions and requiring respondents to rank facts for each case they evaluate. We chose to double the payment compared to the other surveys because trial runs of all four surveys revealed that the structured survey took about 40 minutes to complete, on average, and the others averaged around 20 minutes. To improve data quality, an issue which will be discussed further in section 3.2, we also gave incentives to survey-takers; those who predicted recidivism with an accuracy greater than or equal to 65% would receive double the payment — $6.00 in total for the structured survey, and $3.00 for the others. This payment scheme and amount aligns with previous work that administered decision-making surveys of similar lengths.16-43

We recorded participants’ age bracket, gender, race, highest level of educational attainment, and familiarity with the United States’s criminal justice system; summaries of these demographics are provided in tables 3.1-3.5. Participants were also asked for general reflections on the survey afterward; in a small trial round of the survey, whose other results were not analyzed, these responses were used to clarify question wording and the survey instructions.
2.5 Experiment 1: Noise Notice

This experiment is derived from Kahneman et al.’s idea that a one-time noise audit that tells organizations how much noise exists in their system would help them understand the prevalence of noise and thus eventually reduce it. In this experiment, we inform participants of the typical amount of noise present in recidivism predictions, based on the amount of noise that was present in our control survey results, and then test the effects of this advisement on noise, accuracy, and fairness.

**Hypothesis**

Awareness of the typical level of noise in human judgements in the criminal justice domain will reduce the level of noise in future judgements, where noise is defined as both the level of dissimilarity with which individuals treat similar cases and overall group noise, quantified through metrics like variance or typical agreement among participants. We hypothesize this because having noise explicitly pointed out to them before they are asked for a judgment could make participants more likely to check their own decisions for internal consistency, therefore minimizing noise within one judge’s set of decisions, even if participants cannot know exactly how their decisions compare to others’. We further hypothesize that such awareness of typical noise levels will reduce future noise levels without reducing accuracy metrics (e.g. lowering accuracy or increasing the false positive rate) or measures of fairness. Chapter 3 will define in detail our means of measuring noise, accuracy, and fairness.

**Task**

This experimental survey differed from the control survey only in that, alongside the details of each case, participants were provided a brief notice about the “typical noise levels” that we found in the results of our control survey. The “noise notice,” as it will heretofore be referred to as, read as follows:

In previous iterations of this survey, averaged across all cases, typically no more than [PROPORTION A] percent of participants agreed on a single risk score, and no more than [PROPORTION B] agreed on whether defendants would commit another crime. Individual participants tended to give sim-
ilar cases different scores [PROPORTION C] and different reoffending predictions [PROPORTION D] percent of the time.

Proportion A is defined as the percentage of people who agreed on the risk score that was most commonly assigned. Proportion B is the proportion of people who agreed on whichever recidivism prediction — either that the defendant would not reoffend within two years, or that they would — was chosen the most often for a given case, averaged across cases. Proportion C is the average, across all prior survey participants, proportion of times (out of our 10 manually chosen, “similar” pairs of cases) that participants gave different risk score to the cases in the pair. Similarly, proportion D is the average, across all prior survey participants, proportion of times (out of the 10 pairs of similar cases) that participants gave different recidivism predictions to the pair of cases. These proportions will be defined more formally in section 3.3, as we use these same metrics to quantify the noise in our survey results. This noise notice intentionally does not reveal to participants the contents of prior participants’ decisions, just the aggregate level of agreement, so as to not otherwise bias decisions. We used proportions of agreement as a proxy for noise in this experiment to provide what we hypothesize is a more accessible, easily comprehensible means of visualizing noise compared to standard statistical measures like standard deviation or variance.

For future work, a fully automated version of this potentially noise-reducing technique would likely analyze judges’ decisions as they are submitted and then provide judges with continuously updating snapshots of the organization’s noise levels.

2.6 Experiment 2: Structuring Human Decision-Making

This experiment is based on Kahneman et al.’s hypothesis that providing decision-making processes with more structure could help reduce noise. He postulates that this could occur because asking these questions imposes a more machine-inspired, consistent thinking process on decision-makers.

Hypothesis
Our hypothesis is that imposing more structure and rules on the human-decision making process will reduce participants’ internal noise levels (the dissimilarity of individual participants’ decisions on similar cases) and thus overall group measures of noise, including variance, while improving or at least not harming accuracy and fairness.

**Task**

This task, after presenting the details of each case and defendant, “imposed more structure” on participants’ thinking by asking them to describe in free-form text their decision-making process and to rank the facts most relevant to each case. Only after answering these questions were participants asked to rate defendants’ risk score and predict whether they will reoffend. The text of these structuring questions was as follows:

**Question 1:** What are the key factors influencing your evaluation of this case?

**Question 2:** Rank the following facts of the case by their relevance to your decision:

- Defendant’s age
- Defendant’s race
- The crime they are being charged with
- The severity of the crime
- The defendant’s number of non-juvenile crimes
- The defendant’s number of juvenile misdemeanors
- The defendant’s number of juvenile felonies
- Outside facts you know about the criminal justice system
- Other: please specify
The rest of this experimental survey matches the control survey. For future work, a fully automated version of this strategy that works alongside judges in real time would likely look the same as our survey does; the software would simply also include these structured questions before allowing a judge to input their final decision.

2.7 Experiment 3: Noise Calibration Test

This experiment is also based on Kahneman et al.’s noise audit proposal. However, unlike the Noise Notice experiment, which describes to current participants the noise levels in previous participants’ decisions, in this experiment we aim to give participants a more personalized view of how their decisions compare to previous participants’ and how inconsistent their own decisions are.

Hypothesis

We hypothesize that seeing the results of some of their own decisions compared to previous participants’ decisions will result in participants’ subsequent decisions being more internally consistent (e.g. similar cases will more often be given similar risk scores and similar recidivism predictions) and will either or lower or keep levels of group noise across all participants (e.g. variance, maximum prediction agreement) the same. We hypothesize that accuracy and fairness will simultaneously not be worsened.

Task

This experimental survey differed from the control in that participants were asked to first evaluate — by giving risk ratings and recidivism predictions — a 15-case subset of our full 40 cases, including five pairs of “similar” cases. This subset of cases comprises the “calibration test.” We then quantified the noise and inconsistencies in each participant’s decisions on this subset while they were still in the survey using Javascript embedded in the questions. We then informed participants of their personal noise levels before they evaluated the remainder of the cases. Specifically, we provided the following notice after the first 15 cases:
Your risk scores agreed with [PROPORTION A] percent of previous respondents, on average, and your recidivism prediction agreed with [PROPORTION B] percent of previous respondents. You gave similar cases different scores [PROPORTION C] percent of the time, and you gave similar cases different predictions [PROPORTION D] percent of the time.

Proportion A is defined as the average proportion, among the 15 cases that the participant has rated so far, of previous respondents who provided the same risk score as the current respondent. Proportion B, similarly, is the average proportion of previous respondents who provided the same recidivism prediction as the current survey participant. Proportion C is the proportion of the five “similar” cases that the participant gave different risk scores to, and Proportion D is the proportion of similar cases that the participant gave different reoffending predictions for.

In future work, a fully automated version of this experiment would likely emit a warning whenever it detected that similar cases were given dissimilar judgments, by the same or even different judges, under some stakeholder-chosen definition of similarity.
3.1 Data Overview

300 Mechanical Turk workers participated in our surveys: 75 in each experimental treatment, and 75 more in our control. Experiments were conducted over a two week period at variable times of day.

After excluding low-quality data, a process that will be outlined in Section 3.2, our remaining sample of 241 participants sample self-reported as 51.5% male and 47.7% female, with 0.0% of respondents
declining to give their gender and 0.4% identifying as non-binary or with a third gender. 74.7% of respondents identified as white, 13.3% as African-American or Black, 5.8% as Asian or Asian-American, 5.4% as Hispanic, and 3.7% as Native American. Participants were allowed to select more than one race or ethnicity option. 29.5% were between 18 and 29 years old, 49.0% were 30-44, 14.5% were 45-54, 6.6% were between 55 and 64, and 0.4% were 65 or older.

A majority of participants, 63.9%, reported having earned a bachelor’s degree as their highest level of education. 4.6% earned an associate’s degree as their highest level of education. 25.7% additionally received a master’s degree, and 0.4% reported a professional or graduate degree as their highest level of education. 2.9% reported their highest level of education as a high school degree or the equivalent, and an additional 2.5% completed some college. No respondents reported not having earned a high school degree or the equivalent.

In our surveys’ pre-questionnaire, participants were asked to rate their familiarity with the U.S. criminal justice system. 0.8% responded that they were “not at all familiar”, 11.2% were “slightly familiar”, 31.5% were “somewhat familiar,” 36.9% were “moderately familiar,” and 19.5% responded that they were “extremely familiar.” In aggregate, 87.9% were thus at least somewhat familiar with the system they were being asked to make decisions in.

Tables 3.1-3.5 summarize participant demographics, including a comparison of the set of participants pre- and post-low-quality data exclusion. 16.3% of participants were excluded for having signals of low-quality survey responses; our exclusion process is described in Section 3.2 on data quality.

<table>
<thead>
<tr>
<th></th>
<th>Before Exclusion</th>
<th>After Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>56.3%</td>
<td>51.5%</td>
</tr>
<tr>
<td>Female</td>
<td>43.0%</td>
<td>47.7%</td>
</tr>
<tr>
<td>Non-binary / other gender</td>
<td>0.3%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Declined to say</td>
<td>0.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 3.1: Participants’ gender, before and after low-quality responses were excluded.
### Table 3.2: Participants' self-reported race and ethnicity, before and after low-quality responses were excluded. Percentages do not add to 100%; participants were allowed to select multiple options.

<table>
<thead>
<tr>
<th></th>
<th>Before Exclusion</th>
<th>After Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>African-American / Black</td>
<td>11.7%</td>
<td>13.3%</td>
</tr>
<tr>
<td>Asian-American / Asian</td>
<td>8.0%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>6.0%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Native American or Pacific Islander</td>
<td>5.3%</td>
<td>3.7%</td>
</tr>
<tr>
<td>White</td>
<td>74.8%</td>
<td>74.7%</td>
</tr>
</tbody>
</table>

### Table 3.3: Participants' self-reported age bracket, before and after low-quality responses were excluded.

<table>
<thead>
<tr>
<th>Age Bracket</th>
<th>Before Exclusion</th>
<th>After Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-29</td>
<td>28.0%</td>
<td>29.5%</td>
</tr>
<tr>
<td>30-44</td>
<td>52.3%</td>
<td>49.0%</td>
</tr>
<tr>
<td>45-54</td>
<td>14.0%</td>
<td>14.5%</td>
</tr>
<tr>
<td>55-64</td>
<td>5.3%</td>
<td>6.6%</td>
</tr>
<tr>
<td>65 or older</td>
<td>0.3%</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

### Table 3.4: Participants' self-reported education level, before and after low-quality responses were excluded.

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Before Exclusion</th>
<th>After Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some high school or less</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>High-school degree or equivalent</td>
<td>5.7%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Some college</td>
<td>3.3%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Associate’s degree</td>
<td>4.0%</td>
<td>4.6%</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>61.7%</td>
<td>63.9%</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>25.0%</td>
<td>25.7%</td>
</tr>
<tr>
<td>Graduate or professional degree</td>
<td>0.3%</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

Table 3.2, Table 3.3, and Table 3.4 show the distribution of participants' demographics before and after excluding low-quality responses. The tables highlight changes in percentages for race, ethnicity, age bracket, and education level.
Table 3.5: Participants’ self-reported familiarity with the criminal justice system, before and after low-quality responses were excluded.

<table>
<thead>
<tr>
<th>Familiarity Level</th>
<th>Before Exclusion</th>
<th>After Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all familiar</td>
<td>1.7%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Slightly familiar</td>
<td>12.0%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Somewhat familiar</td>
<td>28.3%</td>
<td>31.5%</td>
</tr>
<tr>
<td>Moderately familiar</td>
<td>37.6%</td>
<td>36.9%</td>
</tr>
<tr>
<td>Extremely familiar</td>
<td>20.3%</td>
<td>19.5%</td>
</tr>
</tbody>
</table>

The gender breakdown of our participants became split more evenly between men and women after our removal process. The proportion of white participants remained roughly the same, but the proportion of Asian or Asian American, Hispanic, and Native American or Pacific Islander participants declined. The proportion of African-American or Black respondents saw the biggest increase.

The age breakdown of participants held fairly stable through the exclusion process. We saw the clearest demographic changes in the breakdown of participants’ highest level of educational attainment and self-assessed familiarity with the U.S. criminal justice system. The proportion of respondents with lower levels of educational attainment — specifically, those without a college degree — decreased by 3.6 percentage points, while the proportion of respondents with a college degree or higher correspondingly increased. In fact, the proportion of respondents in every sub-category of high educational attainment (i.e. associate’s degree, bachelor’s degree, master’s degree, and a higher graduate or professional degree) increased.

Notably, the proportion of respondents who reported being “not at all” familiar with the criminal justice system dropped by more than half via the low-quality data exclusion process. The proportion of participants who were “slightly familiar” dropped as well, albeit less precipitously. Interestingly, the proportion of respondents with the highest levels of familiarity with the criminal justice system (moderate or extreme) also declined slightly. The proportion of participants with a middling level of familiarity, who reported being “somewhat familiar” is what increased by several percentage points.
3.2 **Data Quality**

Concerns have risen since 2018 about the quality of survey data gathered via Amazon’s Mechanical Turk. As such, we monitored several indicators of data quality throughout the pilot and main runs of our surveys. We tracked the amount of time participants took to complete each survey, which was captured by Qualtrics; the failure rate of a preliminary comprehension question that accompanied each survey’s instructions; the failure rate of an attention check question that was randomly displayed among the regular case evaluation and prediction questions; the proportion of participants that gave every case the same risk score; and the proportion of participants that gave every case the same recidivism prediction.

Monitoring these data points did reveal initial data quality issues. In a first pilot round of each of the experiments, in which participants were not awarded any accuracy-based bonuses, up to 20% of participants were completing our surveys in under 5 minutes. Our control, noise notice, and calibration test surveys were expected to take 20 minutes, and our structured survey was expected to take 40 minutes. A completion time of under 5 minutes would indicate a participant was taking fewer than 8 seconds to read and evaluate each case. We decided this was too little time for cases to be evaluated meaningfully. Several participants took fewer than 2 minutes to complete a survey, indicating they were likely responding at random. An additional non-disjoint set of about 20 percent of participants was also giving all the cases the same recidivism prediction, although we have less justification for this being a signal of random, low-quality response. A small proportion of respondents would give each case the same risk score, which we interpreted as a strong signal of the survey not being taken meaningfully.

Our work aims to study the effects of various interventions on relative noise, fairness, and accuracy levels in decision-making, rather than absolute values, meaning that low-quality responses are less harmful in our data than they might be in other research as long as they are present in all experiments. However, we still decided to run our main experiments with accuracy bonuses so as to minimize the number
of low-quality responses, particularly since our rate of low-quality responses varied significantly between experimental treatments in the pilot, non-bonus-incentivized run of our experiments. We decided ahead of time to remove all responses that were completed in under 5 minutes, that failed the attention check, that gave all cases the same recidivism risk score, or that gave all cases the same prediction if doing so was highly correlated with failing the task’s initial comprehension check. We qualified the removal of everyone who gave all cases the same prediction since whether this behavior is indicative of low-quality data or is actually a participant’s true thoughts is somewhat unclear. Failing the comprehension check at a significantly higher rate would indicate that those predicting the same for each case were either not understanding the task at hand or were not meaningfully taking the survey. Removing the responses of everyone who gave all the cases the same prediction also helps unbias our noise analysis: measuring the overall variation in responses could be skewed by significantly different rates of participants predicting yes to every case, which was a pattern far more common than participants predicting no recidivism in every case.

The average time taken to complete each of our surveys was 15 minutes, 16 minutes, 42 minutes, and 14 minutes, respectively, for the control, noise notice, structured, and calibration test experiments. Across all 300 participants in these experiments, 8 took fewer than 5 minutes, 1 failed the attention check, and 4 gave all cases the same score. These participants’ data was removed from analysis. Additionally, we found that in 3 of the experimental groups, the rate of failing the comprehension check was 20 to 30 percentage points higher among participants who gave all cases the same prediction than it was among those who did not. In the fourth experimental group, we found the rates to be roughly the same. Thus, we excluded data from participants who gave all cases the same prediction, since we had essentially received two signals from them of not understanding the task — the far-higher rate of failing the comprehension check and the unanimous predictions.

Before running our experiments, we chose not to blanket exclude all participants who failed the comprehension check since this question was only asked alongside the instructions at the very beginning of
the survey. More concise instructions for the survey accompany every case, so failing the comprehension question at the beginning was itself only a weak signal of misunderstanding the task in our view.

After these removal practices, the sample sizes of our datasets were 52, 61, 65, and 63, respectively, for the control group and the noise notice, structured, and calibration test experiments. Some responses were removed for multiple reasons (e.g. giving all the cases the same risk score and giving all cases the same recidivism prediction, or being completed in under 5 minutes and failing the attention check). Our overall number of participants post-exclusion was thus 241, with 16.3% of respondents who filled out our surveys having been removed.

Tables 3.1-3.5 illustrate the breakdown of participant demographics before and after the low-quality data exclusion process. Trends in the breakdown of participants’ familiarity with the criminal justice validate our removal process — the proportion of participants who self-reported being “not at all familiar” with the U.S. criminal justice system, for example, is halved by our process.

3.3 Noise Analysis

We had hypothesized that each of the three experimental treatments would result in participants’ decisions becoming less internally noisy, meaning that they would more often give similar defendants similar risk scores and recidivism predictions. We further hypothesized that the noise notice and structured experiments would result in less aggregate noise across the group of participants, meaning that people would agree more often and that variance would be lower, and that the calibration test would keep these aggregate measures of noise the same.

We will now define five metrics to analyze changes in noise levels among recidivism predictions. These noise metrics are built conceptually off of the definition of noise provided by Kahneman et al. and additionally match the interpretable noise level explanations we gave to participants in the noise notice and calibration test experiments.
The three measures of group or aggregate noise in our analysis are the Average Max Score Agreement, the Average Max Prediction Agreement, and prediction variance. Group noise refers to variation and disagreement among the decisions of the whole survey group, rather than among the set of decision of an individual decision-maker. The Average Max Score Agreement is the average, across all cases, maximum proportion of respondents who agreed on any given risk score. Similarly, the Average Max Prediction Agreement is the average, across all cases, of the proportion of respondents who gave whichever prediction the majority of respondents gave for that case. Variance is calculated by treating the responses to each case as part of a binomial distribution. Precise equations for our custom Average Max Score Agreement (AMSA) and Average Max Prediction Agreement (AMPA) metrics are as follows:

Let $C = \{c_1, \ldots, c_{40}\}$ be the set of defendant cases that participants are evaluating, of which there are 40 in our experiments, and let $R = \{r_1, r_2, \ldots, r_{|R|}\}$ be the set of participants in a particular experiment. Predictions for a case can be either 1 (the defendant will commit another crime within two years) or 0 (they will not). Let $p_{i,j}$ denote the prediction assigned by the $i$th respondent to the $j$th case. Let $s_{i,j}$ denote the risk score (between 1 and 10, inclusive) assigned by the $i$th respondent to the $j$th case. Then,

$$AMS\text{A} = \frac{1}{|C|} \sum_{j=1}^{|C|} \max_{s^* = 1, \ldots, 10} \left| \left\{ r_i \in R : s_{i,j} = s^* \right\} \right|$$

In this definition, the maximizing statement is simply finding the proportion of respondents who gave the mode risk score for a given case. We then average this proportion across cases to get a metric for the whole survey group that represents the average maximum level of risk score agreement among respondents. Later in this work, we will refer to the AMSA as well as the related Max Score Agreement (MSA), a case-specific metric that quantifies the maximum proportion of respondents who agreed on any risk score for that case. We will refer to the MSA in context of plotting it for all 40 cases to see the spread of values before we focus just on their average, the AMSA.
$AMPA = \frac{1}{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}|} \max_{p^* = 0.1} \left| \left\{ r_i \in R : p_{i,j} = p^* \right\} \right| / |R|

Similarly, in this definition of AMPA, the maximizing statement is finding, for each case, the proportion of respondents who gave the mode recidivism prediction. Conceptually, this is the maximum proportion of respondents that agreed on a recidivism prediction for a given case. We then average this max proportion over all cases, as we did with the AMSA. Again, later in this work, we will also plot the spread of the related Max Prediction Agreement (MPA) metric, which is case-specific. The MPA values across across all cases are what we average to calculate the AMPA.

Higher values of AMSA and AMPA correspond to lower noise levels, since higher agreement levels indicate less variation and inconsistencies between judges.

The two measures of individual noise or internal participant decision consistency we have defined are the Pair Score Difference Rate and the Pair Prediction Difference Rate. Both of these metrics analyze inconsistencies in an individual participant’s judgement on the ten pairs of similar cases we included in our set of 40 cases, where case similarity is defined earlier in our methodology. The Pair Score Difference Rate (PSDR) is the average, across all respondents, proportion of pairs that were assigned different risk scores by the respondent. Similarly, the Pair Prediction Difference Rate is the average, across all respondents, proportion of pairs that were given different recidivism predictions.

Let $a_j$ be the $j$th pair of similar cases; the cases in this pair are $c_j$ and $c_{j+10}$. Let $\mathcal{A}$ be the set of such pairs, such that $|\mathcal{A}| = 10$, since we have 10 pairs of similar cases in our 40-case dataset.

$$PSDR = \frac{1}{|R|} \sum_i \frac{|\{a_{ij} : s_{ij} \neq s_{i,j+10}\}|}{|\mathcal{A}|}$$

The fraction in the above equation refers to the proportion of pairs that respondent $i$ does not give the same risk score to. We then average this proportion across all respondents to see how often participants assign similar cases different scores.
\[
PPDR = \frac{1}{|R|} \sum_{i} \left| \{a_i : p_{i,j} \neq s_{i,j+10}\} \right| / |A|
\]

Similarly, the fraction in the PPDR definition is the proportion of pairs that respondent \(i\) gives different recidivism predictions to, and again we average this proportion across all respondents to see how often participants give similar cases different predictions. Higher values of PSDR and PPDR correspond to higher noise levels, since treating similar cases differently more often indicates a higher level of internal inconsistency for the decision-maker.

The values of these four metrics, as well as the variance of the recidivism predictions, are summarized in Table 3.6 and Figure 3.1. For the calibration test experiment, we report all metrics — the noise values, in this section, and other metrics in later sections — across just the last 25 cases that participants saw after their personalized, calibrated noise notice was presented. As a reminder of this difference from the other experiments, a (25) will be appended to references to this experiment in our results tables. To test for significant differences between our control group and the calibration test experiment, we compared metrics evaluated across all 40 cases in the control group against metrics evaluated across just the 25 post-notice cases in the calibration test group.

<table>
<thead>
<tr>
<th></th>
<th>AMSA</th>
<th>AMPA</th>
<th>PSDR</th>
<th>PPDR</th>
<th>Prediction Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.2135</td>
<td>0.6817</td>
<td>0.7346</td>
<td>0.2827</td>
<td>7.3353</td>
</tr>
<tr>
<td>Noise Notice</td>
<td>0.2180</td>
<td>0.6590</td>
<td>0.7443</td>
<td>0.2869</td>
<td>10.4666</td>
</tr>
<tr>
<td>Structured</td>
<td>0.2081</td>
<td>0.6504</td>
<td>0.7369</td>
<td>0.2677</td>
<td>12.7300</td>
</tr>
<tr>
<td>Calibration (25)</td>
<td>0.2013</td>
<td>0.6463</td>
<td>0.7492</td>
<td>0.2857</td>
<td>11.6907</td>
</tr>
</tbody>
</table>

Table 3.6: Our full set of noise metrics on the control group and in each of the three experiments.

The Pair Score Difference Metric above is a fairly strict metric, as it calculates the proportion of similar cases that were given scores that are not exactly equal. We thus added to our analysis the Relaxed Pair Score Difference Rate (Relaxed PSDR), which calculates the proportion of similar cases that were
given risk scores with an absolute difference greater than 2, averaged across all respondents in a given experiment group. This metric captures actual similarity in defendant treatment, rather than whether defendants are treated exactly the same, which is valuable since the manufactured pairs of similar defendants we included in our cases did not have identical criminal histories. We allowed their priors to be off in count by 1 or 2 across multiple categories of criminal offenses. Table 3.7 contains a comparison of the values of the strict and relaxed metrics.

Figures 3.1 and 3.2 shows boxplots of the spreads of each of these five noise metrics — the four original ones, plus the relaxed PSDR – for each of the four survey groups, as well as plots of the differences in means of each of our metrics between the experimental groups and control group. 95% confidence intervals are included to illustrate the lack of significance in any of these differences: confidence intervals that contain 0 indicate a lack of statistical significance in the difference between the control and experimental group for that metric.

Two-sided t-tests showed no significant difference in any of our noise metrics — either group noise metrics or individual, internal metrics — between the control and any of the experimental treatments. P-values from these significance tests are in table 3.8 for reference. We have therefore found no evidence for our hypotheses that our chosen noise reduction strategies noise levels in human decision-making.

<table>
<thead>
<tr>
<th></th>
<th>Pair Score Diff</th>
<th>Relaxed Pair Score Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.7346</td>
<td>0.2058</td>
</tr>
<tr>
<td>Noise Notice</td>
<td>0.7443</td>
<td>0.1918</td>
</tr>
<tr>
<td>Structured</td>
<td>0.7369</td>
<td>0.2138</td>
</tr>
<tr>
<td>Calibration (25)</td>
<td>0.7492</td>
<td>0.2063</td>
</tr>
</tbody>
</table>

Table 3.7: A comparison of the strict Pair Score Difference Rate and the relaxed Pair Score Difference Rate.
Figure 3.1: Group noise metrics

<table>
<thead>
<tr>
<th></th>
<th>AMSA</th>
<th>AMPA</th>
<th>PSDR</th>
<th>PPDR</th>
<th>Relaxed PSDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise Notice</td>
<td>0.5875</td>
<td>0.3929</td>
<td>0.7547</td>
<td>0.9139</td>
<td>0.6879</td>
</tr>
<tr>
<td>Structured</td>
<td>0.4700</td>
<td>0.1865</td>
<td>0.9417</td>
<td>0.6889</td>
<td>0.8199</td>
</tr>
<tr>
<td>Calibration (25)</td>
<td>0.1603</td>
<td>0.1755</td>
<td>0.6679</td>
<td>0.9414</td>
<td>0.9881</td>
</tr>
</tbody>
</table>

Table 3.8: P-values from two-sided t-tests testing if there is a significant difference in any noise metrics between our experimental groups and control group. None of the above values are significant. P-values are listed under the experimental group being compared to the control.
Figure 3.2: Internal, individual noise metrics
3.4 Accuracy

The goal of our work is to test strategies for lowering overall noise levels in human decision-making, but this goal was desirable in the first place partially for its conceptual role in promoting accuracy and fairness. Thus, we hypothesized initially that our strategies would lower noise levels while also increasing or leaving constant desirable accuracy and fairness metrics. In the following sections, we test the second half of that hypothesis by evaluating the effects of our noise interventions on accuracy levels and group fairness metrics.

We evaluated the effect of the three experimental noise reduction strategies on three different accuracy-related metrics, based on the analysis of similar experiments in related decision-making work. We looked at the average accuracy, average false positive rate (FPR), and average positive predictive value (PPV) across our four groups. Studying the FPR and PPV in addition to raw accuracy allows us to capture nuanced effects and better hypothesize at how participants’ decisions changed qualitatively. These three metrics are defined in terms of the following accuracy primitives:

**Accuracy Primitives:**

1. **TN** (true negative): the number of defendants who did not end up committing another crime within the following two years of the crime they were included in our dataset for, and who were correctly predicted to not commit another crime

2. **TP** (true positive): the number of defendants who did commit another crime within the following two years and who were predicted to do so

3. **FN** (false negative): the number of defendants who did commit another crime within the following two but who were predicted not to do so

4. **FP** (false positive): the number of defendants who did not commit another crime within two years but who were predicted to do so
To calculate the accuracy, FPR, and PPV of an individual survey participant, we calculate the above primitives using that participant’s predictions for all the 40 cases. To calculate the overall accuracy, FPR, and PPV of an experiment, we then average the accuracy, FPR, and PPV of all participants in that experiment.

Our three main accuracy metrics are calculated from the above primitives as follows:

\[
\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}
\]

\[
\text{FPR} = \frac{FP}{FP + TN}
\]

\[
\text{PPV} = \frac{TP}{TP + FP}
\]

To reiterate, our hypotheses with respect to accuracy were that our noise sanitation techniques would either leave constant or improve accuracy, FPR, and PPV. Throughout this section on accuracy, to test these hypotheses, we use two-sided t-tests to reflect the uncertainty of the connection between noise and accuracy and the very real possibility that these experiments could in fact hinder accuracy.

Table 3.9 contains the values for these three metrics across our four groups. Figure 3.3 contains box-plots illustrating the spread of the values of respondents’ accuracy metrics in each of the four survey groups. It also contains plots of the mean difference in these three accuracy metrics between the experimental and control groups. 95% confidence intervals on these difference plots show which differences were statistically significant.

Our results show no significant effect from any of our three noise sanitation techniques on the accuracy or false positive rate of predicting recidivism. Participants in the control group predicted two-year recidivism with 58.65% accuracy. On the other hand, participants in the noise notice experiment, struc-
Table 3.9: The values of our three accuracy metrics across the four survey groups.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>FPR</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.5865</td>
<td>0.5846</td>
<td>0.5881</td>
</tr>
<tr>
<td>Noise Notice</td>
<td>0.5697</td>
<td>0.5770</td>
<td>0.5788</td>
</tr>
<tr>
<td>Structured</td>
<td>0.5831</td>
<td>0.4938</td>
<td>0.5936</td>
</tr>
<tr>
<td>Calibration (25)</td>
<td>0.5752</td>
<td>0.5016</td>
<td>0.4953</td>
</tr>
</tbody>
</table>

tured experiment, and calibration test experiment (over just the last 25 cases) predicted recidivism with 56.97%, 58.31%, and 57.52% accuracy, respectively. Under a two-sided, two-sample t-test, none of these are significant differences, with p-values of 0.43, 0.87, and 0.63, respectively.

Our control group had a false positive rate of 58.46%. The noise notice, structured, and calibration test experiments, which had false positive rates of 57.70%, 49.38%, and 50.16%, respectively, corresponding to two-sided t-test p-values of 0.89, 0.09, and 0.13, respectively.

Our results additionally show no significant effect from our noise notice or structured experimental interventions on the PPV of recidivism prediction. The PPV of the control group was 58.81%, whereas the PPV of the noise notice and structured experiments were 57.88% and 59.36%, resulting in non-significant p-values of 0.69 and 0.83, respectively.

However, we did find that our third experimental technique of including a calibration test and informing participants of their individual noise levels, both internally and compared to previous participants, did have a significant effect on positive predictive value (PPV). Specifically, the calibration test treatment lowered the positive predictive value of participants’ predictions significantly. Our control group had a PPV of 58.81%, whereas the calibration test group had a PPV of 49.33%, which a two-sided t-test shows is a significant difference (p = 0.002).

Table 3.10 contains the p-values resulting from two-sided t-tests evaluating the significance of differences between each of our experimental groups and our control group. Bolded in the table is the one significant result, using a significance level of 0.05.
(a) A boxplot showing the median, quartiles and outliers of accuracy values from all respondents in each survey group.

(b) The average accuracy of each of the experimental groups minus that of the control group, with 95% confidence intervals on the differences.

(c) A boxplot showing the median, quartiles and outliers of false positive rate (FPR) values from all respondents in each survey group.

(d) The average FPR of each of the experimental groups minus that of the control group, with 95% confidence intervals on the differences.

(e) A boxplot showing the median, quartiles and outliers of positive predictive (PPV) values from all respondents in each survey group.

(f) The average PPV of each of the experimental groups minus that of the control group, with 95% confidence intervals on the differences.

Figure 3.3: Accuracy values
Table 3.10: Above are the p-values resulting from two-sided t-tests evaluating the significance of differences in accuracy metrics between each of our experimental groups and our control group. The p-values are listed under the experimental group that the control group is being compared against.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>FPR</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise Notice</td>
<td>0.4320</td>
<td>0.8909</td>
<td>0.6932</td>
</tr>
<tr>
<td>Structured</td>
<td>0.8713</td>
<td>0.0889</td>
<td>0.8311</td>
</tr>
<tr>
<td>Calibration (25)</td>
<td>0.6343</td>
<td>0.1284</td>
<td>0.0020</td>
</tr>
</tbody>
</table>

3.5 Fairness

There are numerous ways in computer science and machine learning literature to assess fairness in decision-making. As introduced earlier, the most used techniques in recent work fall into the categories of individual fairness and group fairness. For a classification algorithm — or, in our case, a human-based decision-making process — to be individually fair, it must treat similar defendants similarly. What this means in terms of our methodology is that a highly individually fair decision-maker would give the defendants that we have defined as similar, similar scores and likely the same recidivism prediction. Since this idea of fairness overlaps significantly with our concept of internal noise — specifically our internal noise metrics of Pair Score Difference Rate (PSDR) and Pair Prediction Difference Rate (PPDR) — in this section we will instead focus on evaluating our experimental groups with respect to group fairness metrics.

There are several categories of group fairness definitions in computer science literature. The three main types are anti-classification, calibration, and classification parity. Anti-classification refers to not using race, gender, or other protected demographic characteristics, or any characteristics that could serve as proxies for them, in a decision-making process. This definition stems from the conception of a fair process as one that does not take into account any characteristics over which a person has no control, which could include the area they grew up in and family history in addition to demographic informa-
As not inferring any information about a defendant’s gender or race when a real-world judge in the criminal justice system makes a decision is logistically impractical — and perhaps even undesirable, for those who seek to actively correct for historical racial injustices in the criminal justice system — we chose to disclose defendants’ race and gender in our surveys and thus cannot use the anti-classification definition of fairness. Furthermore, anti-classification has some fundamental limitations: for example, it precludes the decision-maker from using any culturally relevant information that could actually make a decision-making system more fair. The calibration definition of fairness dictates that, conditioning on predicted risk estimates, actual outcomes should be independent of protected attributes. We chose not to use this definition of fairness to evaluate our results because studying outcome differences among defendants with the same predicted risk score effectively requires a larger sample size than we have. The final category of classification parity, requires for fairness that measures of statistical performance, such as the false positive rate, are similar among groups that differ based on protected characteristics, such as white defendants and Black defendants. This is the definition that we will use to evaluate our experiments’ effects on fairness.

To parallel similar work studying the effects of various interventions on recidivism predictions, such as work by Vaccaro and Green, so as to better compare the results of our decision-making interventions, we will study differences in the accuracy, FPR, and PPV among Black defendants in the dataset and among white defendants in the dataset, for each of our four survey groups. We focused on studying potential differences by race and not gender to parallel prior recidivism prediction work, and we are specifically focusing on only white and Black defendants since they are overwhelmingly the most-represented racial groups in our original Broward County dataset.

As we did for accuracy, we hypothesized that our three experimental noise sanitation techniques would leave fairness metrics the same or improve them in a desirable way, such as by lessening racial differences in the false positive rate.

The mechanics of our significance testing for fairness differed slightly from how we evaluated noise,
fairness, and self-perception metrics. For those metrics, we compared values in the control group against values in each of the experimental groups. To evaluate fairness, however, we will be comparing, within each survey group, metrics such as the false positive rate among white defendants versus among Black defendants. An experiment would thus be deemed to make the recidivism prediction process more fair if, for example, in the control group there was a significant difference between the false positive rate among white defendants and Black defendants but there was not a significant difference in the experimental group, ideally not to the significant detriment of overall accuracy.

In the following tables and results, the superscript of b next to the name of a metric, such as FPR\textsuperscript{b}, refers to that metric among only Black defendants in our cases. Similarly, FPR\textsuperscript{w} refers to the same metric evaluated on only white defendants’ cases. We are using two-sided t-tests to test the difference between these metrics, and then are doing the same for overall accuracy and PPV.

Table 3.11 contains the values of these three metrics among only white and among only Black defendants in our control group and three experimental groups. Figure 3.4 visualizes, for each of the four survey groups, the difference in the mean of each of these three metrics between Black defendants and white defendants, with 95% confidence intervals to illustrate which differences were significant. Intervals that contain 0 as the mean difference were not significant.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Noise Notice</th>
<th>Structured</th>
<th>Calibration (25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (W)</td>
<td>0.5838</td>
<td>0.5703</td>
<td>0.6143</td>
<td>0.5261</td>
</tr>
<tr>
<td>Accuracy (B)</td>
<td>0.6172</td>
<td>0.5925</td>
<td>0.5890</td>
<td>0.6021</td>
</tr>
<tr>
<td>FPR (W)</td>
<td>0.5505</td>
<td>0.5492</td>
<td>0.4442</td>
<td>0.5143</td>
</tr>
<tr>
<td>FPR (B)</td>
<td>0.5962</td>
<td>0.5993</td>
<td>0.5197</td>
<td>0.4940</td>
</tr>
<tr>
<td>PPV (W)</td>
<td>0.5484</td>
<td>0.5359</td>
<td>0.5949</td>
<td>0.3451</td>
</tr>
<tr>
<td>PPV (B)</td>
<td>0.6529</td>
<td>0.6433</td>
<td>0.6396</td>
<td>0.5707</td>
</tr>
</tbody>
</table>

Table 3.11: Above are the accuracy, FPR, and PPV values, segmented to only among white defendants and then only among Black defendants, in each of our survey groups.
(a) The average accuracy of predictions made on Black defendants subtracted by the average accuracy made on white defendants in each of the four survey groups, with 95% confidence intervals on the differences.

(b) The average FPR of predictions made on Black defendants subtracted by the average FPR made on white defendants in each of the four survey groups, with 95% confidence intervals on the differences.

(c) The average PPV of predictions made on Black defendants subtracted by the average PPV made on white defendants in each of the four survey groups, with 95% confidence intervals on the differences.

Figure 3.4: Group fairness results
We found that there was no statistically significant difference between Accuracy$^b$ and Accuracy$^w$ in the control group or in the noise notice and structured experiments, implying fairness. In the control and noise notice groups, Accuracy$^b >$ Accuracy$^w$. In the structured experiment, we observed Accuracy$^b < $ Accuracy$^w$. The disparity in accuracy was larger in the control group than for both the noise notice and the structured experiments, though not to a statistically significant extent. However, there was a statistically significant ($p = 0.02$ under a two-sided t-test) difference in the accuracy with which participants in the calibration test group rated white defendants and Black defendants. In this experiment, Accuracy$^w$ was 52.61% and Accuracy$^b$ was 60.21%, implying that the calibrated noise notice we provided may have actually exacerbated differences between how white and Black defendants were evaluated.

We also found that there was no statistically significant difference between FPR$^b$ and FPR$^w$ values within any of the four survey groups. In the control, noise notice, and structured groups, FPR$^b > $ FPR$^w$, though the difference as actually larger for the noise notice and structured groups than it was for the control group. Only in the calibration test group was FPR$^b < $ FPR$^w$.

Our second significant fairness result was that PPV$^b$ was significantly greater than PPV$^w$ in the control, noise notice, and calibration test groups ($p = 0.0007, 0.0003,$ and $3.66 \times 10^{-8}$, respectively), but there was no significant difference between PPV$^b$ and PPV$^w$ in the structured group, implying this experiment might have resulted in greater fairness, at least using one narrow definition of fairness. Interestingly, this greater similarity in PPV in the structured experiment was caused by PPV increasing among the white defendants, not by any change among the Black defendants, for which the PPV was already noticeably higher than among white defendants. This result implies that the noise sanitation technique of structuring the decision-making process improved participants' ability to identify positive recidivism cases meaningfully among white but not Black defendants. However, participants were already able to identify positive cases among Black defendants at a better rate before any interventions than they were able to among white defendants after this intervention.
<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Noise Notice</th>
<th>Structured</th>
<th>Calibration (25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.2258</td>
<td>0.3982</td>
<td>0.3124</td>
<td><strong>0.0206</strong></td>
</tr>
<tr>
<td>FPR</td>
<td>0.4312</td>
<td>0.3983</td>
<td>0.1728</td>
<td>0.727</td>
</tr>
<tr>
<td>PPV</td>
<td><strong>0.0006</strong></td>
<td><strong>0.0003</strong></td>
<td>0.1798</td>
<td><strong>3.6590 \times 10^{-8}</strong></td>
</tr>
</tbody>
</table>

Table 3.12: Above are the p-values resulting from two-sided t-tests evaluating the significance of the differences between these three fairness metrics on white defendants vs on Black defendants within the same experiment. Bolded are the experiment-metric combinations in which there was a statistically significant difference between predictions for white defendants and Black defendants, using a significance level of 0.05. P-values are listed under the experiment that the control is being compared against.

Table 3.12 contains the p-values resulting from two-sided t-tests evaluating the significance of the differences between these three fairness metrics on white defendants vs on Black defendants within the same experiment. Bolded are the experiment-metric combinations in which there was a statistically significant difference between predictions for white defendants and Black defendants, using a significance level of 0.05.

### 3.6 Self-Perception

In our surveys’ post-questionnaire, presented after participants evaluated our 40 cases, we asked them to rate their confidence in their risk scores and predictions, as well as their self-perceived accuracy and fairness of their evaluations. Accuracy and fairness were not defined either colloquially or technically within the survey, leaving the meaning of these terms to respondents’ discretion. In line with previous work, we aimed via these questions to assess whether, regardless of if our experimental treatments actually resulted in more consistent, more accurate, or fairer recidivism evaluations, participants were more likely under any of the treatments to perceive their evaluations as such. This is an important inquiry to conduct before deploying any potential noise sanitation strategies in the real world because disconnects between judges’ perceptions of their consistency, accuracy, and fairness and their actual consistency, accuracy,
and fairness could dangerously create the illusion of progress and systemic reform while none is truly made, facilitating complacency.

Our “confidence” self-perception question asked respondents to rate their level of agreement with the statement that they were confident in their evaluations. These levels of agreements ranged from “Strongly agree” to “Strongly agree.” In our analysis below, we have corresponded the highest level of agreement with our statements — so, the highest amount of confidence — with a score of 7, and the lowest level of confidence with a score of 1. The other levels of agreement, such as “neither disagree nor agree” or “agree” correspond to the intermediate integer scores of 2 to 6. The fairness and accuracy self-evaluation questions similarly asked participants for their level of agreement with statements that “the decisions I made in this study were fair” and “the decision I made in this study were accurate,” respectively. The agreement options and conversion between options and numerical scores were the same as what we used for the confidence question.

One hypothesis for this inquiry is that the structured experiment would cause participants to feel more confident in their decision and to perceive their decisions as more accurate and fairer, based on prior work studying the effects of more comprehensive, structured deliberation in decision-making.46 Our other hypotheses were that the noise notice and calibration test experiments would cause participants to feel less confident in their evaluations and to rate their own estimated accuracy and fairness as lower, since they were being repeatedly told throughout the experiment the extent to which their decisions differed from others.37

Table 3.13 contains survey participants’ average self-assessed confidence, fairness, and accuracy scores in each of the four groups.

Figure 3.5 contains boxplots summarizing the spread of self-evaluations scores that participants gave for each of confidence, accuracy, and fairness. It also displays the mean difference in these values between the experimental and control groups, with 95% confidence intervals visualized to show which differences were significant.
Figure 3.5: Confidence and perception of accuracy, fairness

(a) A boxplot showing the median, quartiles and outliers of participants’ self-evaluations of their confidence in their predictions, on a scale from 1 to 7 where 7 is the most confident.

(b) The average confidence level of the control group subtracted by the mean confidence of each experimental group, with 95% confidence intervals on the differences.

(c) A boxplot showing the median, quartiles and outliers of participants’ self-evaluations of the accuracy of their recidivism predictions, on a scale from 1 to 7 where 7 is the most accurate.

(d) The average perceived accuracy level of the control group subtracted by the mean perceived accuracy level of each experimental group, with 95% confidence intervals on the differences.

(e) A boxplot showing the median, quartiles and outliers of participants’ self-evaluations of the fairness of their recidivism predictions, on a scale from 1 to 7 where 7 is the most fair.

(f) The average perceived fairness level of the control group subtracted by the mean perceived fairness level of each experimental group, with 95% confidence intervals on the differences.
Contrary to our hypotheses, none of our experiments had a significant effect on participants’ confidence in their recidivism evaluations. On a scale from 1 to 7, where 7 represents the highest confidence level, members of the control group rated their confidence as 6.25 on average. The average confidence scores in the noise notice, structured, and calibration test groups were 6.07, 6.22, and 6.19, respectively. Testing the significance of these difference gave non-significant p-values of 0.31, 0.82, and 0.74.

Similarly, we found no significant effects from any of our experimental treatments on self-assessed accuracy levels. The average self-assessed accuracy score in the control group, again on a scale of 1 to 7 with 7 corresponding to the strongest assertion that their evaluations were accurate, was 6.06. The average scores among the noise notice, structured, and calibration test groups were 5.97, 5.94, and 5.90, respectively. Performing two-sided t-tests on the differences between these scores and the control group’s gave p-values of 0.58, 0.49, and 0.43, respectively.

We did find that the calibration test experiment had a significant effect on respondents’ self-assessed fairness levels. Specifically, these participants, who had seen personalized information about their decision-making noise partway through the survey, perceived their decisions as fairer, contrary to our hypothesis for this experiment. The average fairness score in the control group was 5.62, whereas the average in the calibration test group was 6.10. Using a two-sided t-test, we found that this was a significant increase, with a p-value of 0.026. Both the noise notice and structured experiments also had nearly-significant results. On average, these groups rated the fairness of their own decisions as 6.02 and 6.03, respectively, out of 7. Performing a two-sided t-test of these values against the mean fairness rating of the control

<table>
<thead>
<tr>
<th></th>
<th>Confidence</th>
<th>Accuracy</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>6.2500</td>
<td>6.0577</td>
<td>5.6154</td>
</tr>
<tr>
<td>Noise Notice</td>
<td>6.0656</td>
<td>5.9672</td>
<td>6.0164</td>
</tr>
<tr>
<td>Structured</td>
<td>6.2154</td>
<td>5.9384</td>
<td>6.0308</td>
</tr>
<tr>
<td>Calibration</td>
<td>6.1905</td>
<td>5.9048</td>
<td>6.0952</td>
</tr>
</tbody>
</table>

Table 3.13: Above are the average self-reported confidence, accuracy, and fairness scores among the four groups.
Table 3.14: Above are the p-values from two-sided t-tests evaluating the significance of the difference in self-reported confidence, accuracy, and fairness between the control group and each of the three experimental groups. We use a significance level of 0.05, so significant values under 0.05 are bolded. P-values are listed under the experimental group that the control is being compared against.

<table>
<thead>
<tr>
<th>Group</th>
<th>Confidence</th>
<th>Accuracy</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise Notice</td>
<td>0.3073</td>
<td>0.5797</td>
<td>0.0635</td>
</tr>
<tr>
<td>Structured</td>
<td>0.8201</td>
<td>0.4858</td>
<td>0.0627</td>
</tr>
<tr>
<td>Calibration Test</td>
<td>0.7404</td>
<td>0.4345</td>
<td>0.0255</td>
</tr>
</tbody>
</table>

Each group gives p-values of 0.06 for each.

Table 3.14 contains the p-values resulting from two-sided t-tests evaluating the effect of our experiments on self-assessed confidence, fairness, and accuracy.
4.1 Noise Analysis

None of our three noise sanitation strategies adapted from Kahneman et al. showed any evidence of reducing noise levels. Additionally, out of all of our four survey groups, the control group without any attempted noise-reducing strategies actually had the lowest level of variance in recidivism predictions.

Although our results neither align with our hypotheses on noise levels nor match Kahneman et al.’s
hypotheses on how effective certain general decision hygiene practices would be, we interpret these results as showing the difficulty in lowering noise levels, rather than already the impossibility of it or the broader invalidity of Kahneman et al.’s proposals. Our experimental designs represent merely one imagined way of adapting Kahneman’s ideas and automating them to operate in settings where many judges are making many similar decisions.

Our results do align with literature demonstrating the lack of impact that notices or disclaimers tend to have on human judges’ decisions in repeated, case-base decision making settings like the ones used in our experiments. Our work also interestingly parallels prior results that show that various different means of explaining algorithmic risk assessments had little to no effect on the final decisions human judges made when consulting said risk assessments. We similarly found that our noise notices — which, essentially, are an attempt to in some part explain the human decision-making process in the criminal justice system via its variability and lack of consensus — had no effect on judges’ final decisions in terms of noise or, as we will discuss further later, accuracy. Our result that providing noise notices — either general or personalized, calibrated ones — does not move noise levels aligns with results from psychology literature that say that neither warning decision-makers of potential biases nor providing personal feedback on biases significantly helps unbias decision-making processes.

One limitation of this work is the lack of nuance in the content of our noise notices. Both the noise notice and the calibration test experiments used very similar noise notices — the wording and content was essentially identical; the main difference resided in what data the presented metrics were calculated from and when the notice was presented to participants. As such, we did not get to evaluate the effects of different kinds of notices to test hypotheses such as whether highly technical explanations of noise or layman’s explanations might be more effective, or whether the timing or repetition of the notice might matter, for example. Admittedly, prior work has found that different styles of explaining recidivism algorithms did not affect subsequent decisions, but such results do not necessarily a priori hold in contexts where the explanation is about a process other than an algorithm.
Another limitation of this work is our narrow interpretation of Kahneman et al.’s conceptions of noise and their strategies for noise reduction and general decision hygiene. Our methodology focused on testing whether highly scaled, automated implementations of Kahneman et al.’s proposals affected noise levels. For example, Kahneman et al. theorized that having organizations perform a one-time noise audit to quantify current levels of noise in their decision-making process, inform stakeholders in the decision-making organization of these levels, and stress the importance and potential effects of noise was an important first step in overhauling the decision hygiene practices of a decision-making body. We interpreted their emphasis on the strategic importance of these noise audits as suggestive of the general potential efficacy of informing decision-makers about noise levels before they make decisions. But perhaps this step of informing stakeholders of noise levels is more relevant as a priming step to motivate organizers to want to try out actual strategies to reduce noise rather than a potentially effective strategy itself. Additionally, perhaps adding automation into these strategies, which our work is predicated on, rather than having them be a more informal, interactive process weakens the potential efficacy of the strategy’s core concept. Thus, taking a broader view of the decision-making process to identify more points where interventions could help lower noise might be useful. Recent work has found, for example, that using bipartite matching algorithms to select which judges even address which cases can improve the accuracy and fairness of decisions. ⁴⁶ Perhaps a similarly situated intervention could address noise levels as well.

A key limitation of methodologies like ours in general, especially when it pertains to decision-making in criminal justice, should anyone want to apply our results as actual insights about the system, is data quality. Firstly, our participants are neither judges nor experts nor particularly knowledgeable in criminal justice, the field we were asking them to make decisions in. Perhaps the decision-making noise produced by non-expert participants is more difficult to overcome than decision noise made by experts, experts who might be poised to better understand noise notices like ours, for example. And secondly, as noted in our Data Quality section, we saw issues with the quality of Mechanical Turk responses that parallel
concerns raised in prior work. Though we did exclude data that showed signals of being low-quality, having in the first place entries with impossibly quick response times, high rates of failed comprehension checks, and perfunctory answers to our open ended questions suggest that even if we cannot test actual judges for the effects of noise reducing strategies, Mechanical Turk at least might no longer be as useful a platform for doing so as it used to be.

4.2 Accuracy

None of our interventions in the decision-making process had any significant effect on participants’ accuracy or false positive rate in predicting two-year recidivism. Accuracy in each experiment hovered around 56-58%. Literature presents contrasting results on whether other forms of notices — for example, presenting the results of an algorithmic assessment, with and without explanation of the algorithm — improve accuracy; some found it did, and others did not. Literature does find that human decision-makers’ accuracy rests solidly below that of risk assessment algorithms like COMPAS, which means our work does not identify a strategy that can reliably outperform algorithmic risk assessment.

We did find that the calibration test experiment resulted in a significantly lower PPV than we had in the control group. This result means that the among the defendants who participants predicted would commit another crime, the proportion of people who actually later did was significantly lower. Though we did not focus on studying the false negative rate (FNR), it was correspondingly higher in the calibration test group (0.3095) than in the control group (0.2430). Furthermore, though this was not a statistically significant difference, the FPR in the calibration group was about 8 percentage points lower than in the control group. These trends occurring together likely signify a generally lower rate of predicting positive recidivism in the calibration test group more than they signify anything about PPV individually. That itself could be a meaningful result, however, depending on what the overall decision-making system values. For example, if the system’s stakeholders see falsely predicting someone will commit a crime
as a more costly error than predicting that someone will not even when they go on to do so, then the 
calibration test strategy could be a valuable solution in potentially engendering caution before a positive 
prediction is made. Neither of the other experiments affected the PPV of recidivism prediction.

As an aside, one factor that might have anchored the accuracy rates of our participants is that our 
Mechanical Turk incentive scheme doubled the task payment to any survey respondent who predicted 
recidivism with greater than 65% accuracy; participants were made aware of this threshold. However, 
this incentive model was used across all four surveys; as such, we expect it had no effect on accuracy dif-
ferences and thus no effect on our analysis.

4.3 Fairness

To evaluate fairness, we used group definitions, specifically by analyzing whether there were significant 
differences between the accuracy, FPR, and PPV of participants’ predictions on white versus Black de-
fendants in each of our three experimental surveys, and then by comparing those fairness results to that 
of our control group. In none of the four surveys sent out was there a significant difference in FPR on 
white and Black defendants. And only in the calibration test experiment was there a statistically signif-
icant difference between accuracy on white and Black defendants; the accuracy on Black defendants 
was higher. This result alone could potentially implicate the calibration test intervention as causing un-
fairness, and specifically as treating white defendants unfairly, since it caused participants to rate white 
defendants’ recidivism with relatively lower accuracy, but the fairness picture becomes more complicated 
when we look at the actual PPV and FPR values for white and Black defendants in this experiment. In 
the calibration test survey, the PPV on white defendants is substantially lower (to a statistically signif-
icant extent) than it is for Black defendants and also noticeably lower than the PPV on white defen-
dants for any other experiment. The FPR is also lower on white defendants in the calibration experiment 
than it is in the control group. Together, these results imply an overall lower rate of positive predictions
among white defendants. We do see an even larger drop in the FPR among Black defendants from the control group to the calibration experiment, but we do not see the same precipitous decline in PPV. Thus, participants in our calibration test were able to pinpoint positive recidivism cases among Black defendants more precisely, and had greater general accuracy than they did on white defendants, but this is at least partially caused by or associated with a greater false negative rate among white defendants (0.3730) than among Black defendants (0.2980), which then could be interpreted as a sign of unfairness against Black defendants.

Of course, it is impossible without a perfect predictor for FPR, PPV, and FNR to be equal among two different groups with different base rates of recidivism, which is part of why we only study closely the first two values. This essentially means that the direction of the unfairness caused by the calibration test experiment (whether it is unfair to Black or white defendants) is unclear and again depends on the decision-making system’s values about which type of errors are the least desirable. However, the calibration test evidently does cause more unfairness when fairness is described through unequal group metrics on different sub-groups, regardless of the ambiguity of who is being treated unfairly and exactly how. This unfairness result matches what Vaccaro found — that human judges seeing the results of a risk assessment algorithm (specifically the COMAPS algorithm) before making their prediction did not affect overall accuracy but did create unfairness by favoring white defendants when analyzing fairness metrics.

On the other hand, we saw evidence that our structured experiment improved fairness by lowering the discrepancy in how Black versus white defendants are treated. In the control group and noise notice and calibration test experiments, we saw an extremely statistically significant difference in the PPV on Black and white defendants. The structured experiment increased the PPV for white defendants to the point where the difference between Black and white defendants was no longer significant.
4.4 Self-Perception

Our main reason for testing participants’ confidence and perception of their own accuracy and fairness was to check if any of our experiments, somewhat dangerously, gave participants the illusion of having made better decisions when the results said they had actually not. We found that participants in the calibration test experiment felt that their decisions were fairer than participants in the control group did. This was a statistically significant result. However, our results showed that the calibration test actually caused more unfairness than was present in the control setting; participants perceiving themselves as fairer, then, is concerning.

Participants in the noise notice and structured experiments also perceived their decisions as fairer, on average, than those in the control group. These were both nearly statistically significant results (p = 0.06 for both), therefore still raising concerns that our experiments might have helped give participants a sense of being more fair than they were. We did actually find that the structured experiment improved fairness by improving the PPV on white defendants, making it more similar to that on Black defendants; that was our only result of having improved fairness. However, the noise notice experiment did not improve fairness outcomes, at least as quantified by any of our chosen metrics, making these participants’ greater perception of their fairness also concerning.

Interestingly, participants rated their accuracy much more highly than was true in reality. The average self-perceived accuracy rating was between 5.9 and 6.1 out of 7 for all four groups; as a percentage, this falls between 84% and 87%. However, actually prediction accuracy rates fell between 56% and 59% for all survey groups.

None of the experiments had a significant effect on participants’ self-evaluation of accuracy or confidence; the latter lack of effect aligns with prior work that found that neither seeing a COMPAS score nor a COMPAS score and a disclaimer about the algorithm before making predictions significantly affected participants’ confidence in their evaluations.16
4.5 Future Work

To address the limitation of a lack of variation in our noise notices, future work could try a range of them, specifically by varying the noise metrics presented in the survey and the level of technicality with which they are explained. Additionally, we could attempt to automate others of Kahneman et al.’s proposals for reducing noise levels. We could also automate the same proposals but in different ways, since our experiments were just one interpretation of their ideas, or we could test the strategies without automating them, in case that is what is interfering with their efficacy.

If future work does identify strategies for reducing noise in decision-making that can be automated, we could build live noise auditing software to check the human decision-making process. One example of how such software could work is as follows:

1. Depending on the decision-making context, decide on metrics for “noise” (e.g. an individual judge’s deviations from their own judgments, deviation from the group’s typical judgements, pure variance and standard deviation, etc.), which should be done in collaboration with the major stakeholders of the decision-making system.

2. Decide on undesirable “extraneous” factors that we do not want affecting judgment (time of day, number of cases evaluated so far that day, judge fatigue level, etc.). Again, this should be worked on with the system’s stakeholders.

3. Build software that, in real time, when a judge is deciding, provides noise warnings calculated from other judges’ data and this individual judge’s past decisions (e.g. you typically disagree with X% of judges; the variance in your cases is X), or that automate other strategies identified in the future as reducing noise.

4. After the decision is made, the software would provide another noise warning if the judge’s decision deviated significantly from a previous case that was extremely similar, flagging as well if...
only the extraneous factors were different (e.g. the similar cases that were decided differently were made at different times of day). Ideally, judges could choose to resubmit a decision or to have another decision-maker added to the judgment loop. The criteria for similarity in cases would also be decided by the main stakeholders of the system.

Additionally, even if finding noise-reducing strategies that can be automated is unsuccessful, we could just build software that automates steps 2 and 4 above, since we know from prior work that factors like break times and time of day can impact decisions when they should not. We could further test whether this intervention causes judges to change their decisions to be fairer or more accurate, or to contribute less to overall noise levels.

Future work could also try to lower noise in the decision-making process by attacking parts of it outside of when the judge is making their final decision, especially if such work if focused on just one decision-making domain, like bail decisions. This work could incorporate other of Kahneman’s ideas, or move beyond that work entirely. For example, as Kahneman hypothesizes would be useful, a future study could have participants collaborate to submit a single risk score and recidivism prediction to see if group work lowers noise levels without sacrificing — or while even improving — accuracy and fairness.

One limitation to our analysis of self-perceived confidence, accuracy, and fairness is evident after looking at the raw scores participants gave. Across all three question categories, scores similarly hover around 6; the quality of data when participants are asked to rate properties that are undefined and could be conceptually overlapping is unclear. Having participants rate their confidence in-survey, as they submit their predictions for each case, might provide more sophisticated insight, for example. Additionally, future work might seek more detailed qualitative insight from participants on what they thought of each of the noise reduction strategies and what they thought of their own decision-making processes. The free-text feedback we received through Mechanical Turk was typically short and cursory, with many respondents submitting a simple “good.”
Finally, a major direction for future research based on our work is to replicate our methodology on actual judges or on people with more expertise in the criminal justice system — or whichever field the example cases are set in — such as law students or undergraduates majoring in fields related to criminal justice.

Ultimately, though we do firmly believe that future work should continue investigating strategies that could reduce noise levels in human decision-making, there always remains the possibility that noise — and perhaps even noticeable, consequential amounts of it — is simply unavoidable for human decision-makers. Perhaps this differentiator of humans and machines is fundamental to what makes human decision-makers human; perhaps it is even valuable.

This is certainly the viewpoint of some, though of course the premise of our work was the idea by Kahneman et al. that it can and should be minimized. Some rationale for the opposite viewpoint, however, is that noise — especially group noise metrics — can simply be a signal of a valid, desirable diversity of perspectives and thought. Noise in individual noise metrics could be a sign of healthy growth and learning in the decision-maker. Some argue that eliminating noise could accidentally eliminate actually useful information or data that we simply have not been able to understand yet. If Kahneman et al.’s strategies to reduce noise continue to not pan out in practice, or else if an organization consciously decides it does not want to lower noise, then future work in the spirit of this thesis could turn toward taking more strategies and proposals like Kahneman’s, either automated or not, and aiming to simply improve accuracy or fairness. Regardless, the pursuit of techniques other than predictive machine learning models to strengthen subjective decision-making processes is important.
In this thesis, we aimed to test whether automated versions of strategies that Kahneman et al. proposed in their book *Noise* to reduce noise in human decision-making actually did just that in practice. Subjective decision-making is ubiquitous in fields such as criminal justice, banking, and academia, with typically extremely high stakes. But humans are, quite naturally, prone to decision fatigue, random errors, and inconsistencies over time in their decision-making; there are often even patterns in seemingly random noise, such when a person is making decisions subconsciously influenced factors unrelated to
the task at hand, like how long they have worked without a break. We hoped to alleviate the impact of such factors and reduce noise in general to support more consistent, fairer, and accurate better decision-making processes.

Thought Kahneman et al. describe in their book several “decision hygiene” strategies to reduce noise levels in a system, we focused on adapting two of them into three separate experiments. Their first strategy is to perform a one-time noise audit of a system in which they present a set of situations or cases typical of the system’s domain — e.g. loan applicant profiles to approve or deny — to many decision-makers and then measure the noise of the group’s decisions, such as the standard deviation of the set of assigned loan scores. As a first step to later reforming the system, they then propose informing stakeholders of these noise levels to highlight the importance and prevalence of noise. A second noise-reducing strategy they propose is breaking down complex judgments into smaller pieces.

Our general methodology involved deploying surveys on Amazon Mechanical Turk that presented to respondents 40 criminal defendant profiles pulled from a real dataset and asked them to give each defendant a risk score and predict whether they would commit another crime in two years; this is what our control group of participants saw. We adapted the noise audit strategy from Kahneman et al. into two experiments in the form of variations on this control survey: the noise notice experiment and calibration test experiment. In the noise notice experiment, alongside the description of each defendant profile, or case, we summarized the noise levels detected among decisions of previous survey respondents. In the calibration test experiment, we had survey participants submit predictions for 15 cases without seeing any information about noise. Then, we calculated the noise within the predictions that that individual survey participant had already submitted, as well as the level of disagreement between that participant’s decisions and previous participants’ decisions. We then presented participants with a noise notice containing this information, similar in structure and content to the one presented in the noise notice experiment, and had them evaluate the final 25 cases. We adapted Kahneman et al.’s idea of breaking down complex judgments into smaller steps into a third experiment by asking participants
to rank the relevance of certain facts — such as the defendant’s age, gender, or history of offences — to their decision-making process and then to explain via free form text their thinking before they evaluated defendants.

We found no evidence of any of our experiments reducing noise under any noise metric we studied. Our noise metrics were separated into two categories: group noise metrics and individual, internal noise metrics. Each of these metrics were not only used to analyze our results but were also the metrics presented to survey participants in the noise notice and calibration test experiments. Additionally, none of our experiments improved participants’ accuracy or lowered their false positive rates in predicting two-year recidivism. We did find, however, that the calibration test experiment lowered the positive predictive value (PPV) of predictions, contradicting our hypothesis that our experiments would at least not worsen any accuracy metrics. This is a mixed result, since though participants became less adept at identifying true positive cases among all the cases they identified as positive, they also generally predicted far fewer positive cases, which might be a more meaningful trend if the decision-making system values an abundance of caution.

Our experiments’ effects on fairness were similarly mixed: the calibration test experiment exacerbated accuracy differences between Black defendants and white defendants and simultaneously maintained from the control group a large difference in PPV between Black and white defendants. Though the accuracy and PPV on Black defendants was higher than on white defendants in this experimental group, it is unclear if this is a sign of unfairness in favor of white or Black defendants. The structured experiment, however, clearly reduced racial disparities in PPV by increasing the PPV value for white defendants, therefore improving fairness under the group definition. Finally, we asked participants at the end of each survey to rate their confidence in their evaluations and to assess their own fairness and accuracy levels. We found that participants in all experiments perceived their decisions as fairer — this result was significant for the calibration test group and nearly significant for the other two — even though only the structured experiment results actually were. Thus, the calibration test, and perhaps the noise notice ex-
experiment too, might have given participants a false sense of fairer evaluation. None of the experiments affected participants’ confidence or self-assessments of accuracy.

Although none of our experiments successfully lowered noise, we do believe further work should go into this line of inquiry. Our experiments represent, firstly, merely one set of interpretations of Kahneman’s strategies and how to automate them, and secondly, only a subset of his original slew of ideas. His ideas are complex and promising; we believe future work could try to automate more of his strategies, test out different interpretations of them, or simply test the core strategies with no software or angling towards eventual automation involved. For example, his suggestion to have two people or a group jointly decide on a risk score and a prediction is one of the most promising paths forward. We could even simply try variations on our existing experiments, making the noise notices more or less technical or explaining the desirability of low noise and its role in consistent, fair decision-making further. We could also make interventions earlier in the decision-making process, since our current efforts focus on how to reduce noise only right as a final decision is made.
A

Survey Excerpts

In this appendix we include screenshots of our surveys, highlighting in particular what our experimental interventions looked like in practice and differences between the control and experimental surveys.
What is the purpose of this research?
We are studying how decisions are made and perceived in the criminal justice field and quantifying various properties of the decisions that judges make. Specifically, we are interested in understanding the tendency of criminals to reoffend and how people perceive the risk of criminals reoffending.

What can I expect if I take part in this research?
We expect that this study should take you about twenty to thirty minutes to complete. First, after giving your consent to participate, you will be asked some demographic questions and to rate your familiarity with the criminal justice system. You will then be asked to evaluate information about 40 individuals charged with criminal offenses. You will see the defendant's age, race, gender, and previous criminal history. You may see supplementary information about the criminal justice system to help guide your decision. You will then assign the person a risk score and predict if they will commit another crime within the next 2 years. You will then be asked some concluding questions reflecting on your experience with the study.

What should I know about a research study?
- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
- You can ask all the questions you want before you decide.

You may not be told everything or may be misled
As part of this research design, you may not be told or may be misled about the purpose or procedures of this research beforehand. However, the purpose or procedures of the research will be disclosed to you following your participation.

Who can I talk to?
If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research team at reoffendingstudymturk@gmail.com.

By clicking the consent button below, you acknowledge that you are 18 or older and that you are aware of the information above.

<table>
<thead>
<tr>
<th>I consent</th>
</tr>
</thead>
<tbody>
<tr>
<td>I do not consent</td>
</tr>
</tbody>
</table>

Figure A.1: Survey consent form (constant across experiments)
(a) Gender and race demographic questions.

What is your gender?

- Male
- Female
- Non-binary / third gender
- Prefer not to say

Which one of these groups do you identify with? (You may select multiple)

- African American or Black
- Asian
- Hispanic
- Native American
- White

(b) Criminal justice familiarity questions

What is your level of familiarity with the criminal justice system in the United States?

- Not at all familiar
- Slightly familiar
- Somewhat familiar
- Moderately familiar
- Extremely familiar

Figure A.2: Opening demographics questions, part 1
What age are you?

- 18-29
- 30-44
- 45-54
- 55-64
- 65+

What is your highest level of educational attainment?

- Some high school or less
- High school degree or equivalent
- Some college
- Associates degree
- Bachelor's degree
- Master's degree
- Professional degree beyond bachelor's and/or doctorate degree

(a) Age and education level.

Figure A.3: Opening demographics questions, part 2
Instructions for your Task

You will be asked to evaluate information about an individual charged with a criminal offense in the United States between 2013 and 2014.

You will see the defendant’s age, race, gender, and previous criminal history. You will then assign the person a risk score (their risk of committing another crime) and predict if they will commit another crime within 2 years.

There will be 40 of these questions. You will be compensated $1.50 for completing the survey in full.

If you predict crime with at least 65% accuracy, you will be awarded a bonus payment of $1.50. As such, please answer the following questions to the best of your ability.

I have read these instructions and am ready to begin the study.

(a) Instructions for the control survey

Instructions for your Task

You will be asked to evaluate information about an individual charged with a criminal offense in the United States between 2013 and 2014.

You will see the defendant’s age, race, gender, and previous criminal history. You will also see a notice with some information about the variation in how previous participants have rated defendants like these. This notice will be the same for all questions.

You will then assign the person a risk score (their risk of committing another crime) and predict if they will commit another crime within 2 years.

There will be 40 of these questions. You will be compensated $1.50 for completing the survey in full.

If you predict crime with at least 65% accuracy, you will be awarded a bonus payment of $1.50. As such, please answer the following questions to the best of your ability.

I have read these instructions and am ready to begin the study.

(b) Instructions for the noise notice experiment survey

Figure A.4: Task instructions, control and noise notice
Instructions for your Task

You will be asked to evaluate information about an individual charged with a criminal offense in the United States between 2013 and 2014.

You will see the defendant's age, race, gender, and previous criminal history. You will then assign the person a risk score (their risk of committing another crime) and predict if they will commit another crime within 2 years. You will also be asked an open-ended question about how you reached your decision and to rank the relevance of the facts you used to make your decision.

There will be 40 of these questions. You will be compensated $3.00 for completing the survey in full.

If you predict crime with at least 65% accuracy, you will be awarded a bonus payment of $3.00. As such, please answer the following questions to the best of your ability.

I have read these instructions and am ready to begin the study.

(a) Instructions for the structured experiment survey

Instructions for your Task

You will be asked to evaluate information about an individual charged with a criminal offense. The data you will view pertains to crimes committed in the United States in 2013 and 2014.

You will see the defendant's age, race, gender, and previous criminal history. You will then assign the person a risk score (their risk of committing another crime) and predict if they will commit another crime within 2 years.

There will be 40 of these questions. You will see 15 of them, then some information about how your results so far compare to previous responses, and then the remaining 25 questions.

You will be compensated $1.50 for completing the survey in full.

If you predict crime with at least 65% accuracy, you will be awarded a bonus payment of $1.50. As such, please answer the following questions to the best of your ability.

I have read these instructions and am ready to begin the study.

(b) Instructions for the calibration test experiment survey

Figure A.5: Task instructions, structured and calibration test
What are you being asked to predict in this study?

- The length of someone's criminal sentence
- Future crime
- Future criminal justice policy
- Defendants' bail amounts

(a) The comprehension check question occurred immediately after task instructions were given. Please select "6" for this question to show you are paying attention.

(b) This attention check question appeared randomly within the survey and an incorrect answer caused immediate data removal.

Figure A.6: Data quality check questions
The defendant is an African-American female aged 26. They have been charged with: possession of Pyrrolidinovalerophenone (colloquially known as flakka). This crime is classified as a felony. They have been convicted of 5 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

Please indicate the likelihood that this person will commit another crime by assigning them a risk score (risk of committing another crime). 1 to 4 are low risk, 5 to 7 are medium risk, and 8 to 10 are high risk.

Select risk score:

Do you think this person will commit another crime within two years?

Yes

No

Figure A.7: Example case (control and calibration test experiments)
The defendant is a Caucasian male aged 28. They have been charged with: battery. This crime is classified as a misdemeanor. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

In previous iterations of this survey, averaged across all cases, typically no more than 21% of participants agreed on a single risk score, and no more than 68% of people agreed on whether defendants would commit another crime. Individual participants tended to give similar cases different scores 73% of the time and different reoffending predictions 28% of the time.

Please indicate the likelihood that this person will commit another crime by assigning them a risk score (risk of committing another crime). 1 to 4 are low risk, 5 to 7 are medium risk, and 8 to 10 are high risk.

---

Do you think this person will commit another crime within two years?

- Yes
- No
The defendant is an African-American male aged 41. They have been charged with: aggravated battery. This crime is classified as a felony. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

What are the key factors influencing your evaluation of this case?

Rank the following facts of the case by their relevance to your decision:

- Defendant's age
- Defendant's race
- Defendant's gender
- The crime they are being charged with
- The severity of the crime
- The defendant's number of non-juvenile crimes
- The defendant's number of juvenile misdemeanors
- The defendant's number of juvenile felonies
- Outside facts you know about the criminal justice system
- Other: please specify
On average, so far, each of your risk scores agreed with 12 percent of previous respondents, and each of your recidivism predictions agreed with 58 percent of previous respondents. You gave similar cases different scores 40 percent of the time, and you gave similar cases different predictions 0 percent of the time.

You will now be asked to evaluate the remaining 25 cases.

I have read this notice.

Figure A.10: Example mid-survey, calibrated noise notice (calibration test experiment)
(a) Participants self-rated their confidence in their predictions

> I am confident in the decisions I made during this study.

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Somewhat agree</th>
<th>Neither agree nor disagree</th>
<th>Somewhat disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

(b) Participants self-rated the accuracy of their predictions

> The decisions I made in this study were accurate.

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Agree</th>
<th>Somewhat agree</th>
<th>Neither agree nor disagree</th>
<th>Somewhat disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
</table>

*Figure A.11: Participant self-evaluation questions, part 1*
The decisions I made in this study were fair.

<table>
<thead>
<tr>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
</tr>
<tr>
<td>Somewhat agree</td>
</tr>
<tr>
<td>Neither agree nor disagree</td>
</tr>
<tr>
<td>Somewhat disagree</td>
</tr>
<tr>
<td>Disagree</td>
</tr>
<tr>
<td>Strongly Disagree</td>
</tr>
</tbody>
</table>

*Figure A.12: Participant self-evaluation questions, part 2*
References


