

# Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics

Debjani Saha<sup>1</sup> Candice Schumann<sup>1</sup> Duncan C. McElfresh<sup>1</sup> John P. Dickerson<sup>1</sup> Michelle L. Mazurek<sup>1</sup>  
Michael Carl Tschantz<sup>2</sup>

## Abstract

Bias in machine learning has manifested injustice in several areas, such as medicine, hiring, and criminal justice. In response, computer scientists have developed myriad definitions of *fairness* to correct this bias in fielded algorithms. While some definitions are based on established legal and ethical norms, others are largely mathematical. It is unclear whether the general public agrees with these fairness definitions, and perhaps more importantly, whether they *understand* these definitions. We take initial steps toward bridging this gap between ML researchers and the public, by addressing the question: *does a lay audience understand a basic definition of ML fairness?* We develop a metric to measure comprehension of three such definitions—demographic parity, equal opportunity, and equalized odds. We evaluate this metric using an online survey, and investigate the relationship between comprehension and sentiment, demographics, and the definition itself.

## 1. Introduction

Research into algorithmic fairness has grown in both importance and volume over the past few years, driven in part by the emergence of a grassroots Fairness, Accountability, Transparency, and Ethics (FATE) in Machine Learning (ML) community. Different metrics and approaches to algorithmic fairness have been proposed, many of which are based on prior legal and philosophical concepts, such as disparate impact and disparate treatment (Feldman et al., 2015; Chouldechova, 2017; Binns, 2017). However, definitions of ML fairness do not always fit well within pre-existing legal and moral frameworks. The rapid expansion of this field makes it difficult for professionals to keep up, let alone the general public. Furthermore, misinformation about notions of fairness can have significant legal implications.<sup>1</sup>

<sup>1</sup>University of Maryland, College Park, MD <sup>2</sup>ICSI, Berkeley, CA. Correspondence to: Michelle L. Mazurek <mmazurek@cs.umd.edu>.

<sup>1</sup><https://www.cato.org/blog/misleading-veritas-accusation-google-bias-could-result-bad-law>

Computer scientists have largely focused on developing mathematical notions of fairness, and incorporating them into ML systems. A much smaller collection of studies have measured public perception of bias and (un)fairness in algorithmic decision-making. However, as both the academic community and society in general continue to discuss issues of ML fairness, it remains unclear how to ensure that non-experts can understand various mathematical definitions of fairness sufficiently to provide opinions and critiques.

**Our Contributions.** We take a step toward addressing this issue by studying peoples’ comprehension and perceptions of three definitions of ML fairness: *demographic parity*, *equal opportunity*, and *equalized odds* (Hardt et al., 2016). Specifically, we address the following research questions:

**RQ1** When provided with an explanation intended for a non-technical audience, do non-experts comprehend each definition and its implications?

**RQ2** Do demographics play a role in comprehension?

**RQ3** How are comprehension and sentiment related?

**RQ4** How do the different definitions compare in terms of comprehension?

We developed two online surveys to address these research questions. We presented participants with a simplified decision-making scenario and a accompanied *fairness rule* expressed in the scenario’s context. We asked questions related to the participants’ comprehension of and sentiment toward this rule. Tallying the number of correct responses to the comprehension questions gives us a *comprehension score* for each participant. In Study-1, we found that this comprehension score is a consistent and reliable indicator of understanding demographic parity.

Then, in Study-2, we used a similar approach to compare comprehension among all three definitions of interest. We find that 1) education is a significant predictor of rule understanding, 2) the counterintuitive definition of Equal Opportunity with False Negative Rate was significantly harder to understand than other definitions, and 3) participants with low comprehension scores tended to express less negative sentiment toward the fairness rule.

## 2. Related Work

In response to many instances of bias in fielded artificial intelligence (AI) and machine learning (ML) systems, ML fairness has received significant attention from the computer-science community. Notable examples include gender bias in job-related ads (Datta et al., 2015), racial bias in evaluating names on resumes (Caliskan et al., 2017), and racial bias in predicting criminal recidivism (Angwin et al., 2016).

To correct biased behavior, researchers have proposed several mathematical and algorithmic notions of fairness.

Most algorithmic fairness definitions found in literature are motivated by the philosophical notion of individual fairness (e.g., see (Rawls, 1971)), and legal definitions of disparate impact/treatment (e.g., see (Barocas & Selbst, 2016)). Several ML-specific definitions of fairness have been proposed which claim to uphold these philosophical and legal concepts. These definitions of “ML fairness” fall loosely into three categories (for a review, see (Chouldechova & Roth, 2018)). *Statistical Parity* posits that in a *fair* outcome, individuals from different protected groups have the same chance of receiving a positive (or negative) outcome. Similarly, *Predictive Parity* (Hardt et al., 2016) asserts that the predictive accuracy should be similar across different protected groups—often measured by the false positive rate (FPR) or false negative rate (FNR) in binary classification settings. Myriad other definitions have been proposed, based on concepts such as calibration (Pleiss et al., 2017) and causality (Kusner et al., 2017). Of course, all of these definitions make limiting assumptions; no concept of fairness is perfect (Hardt et al., 2016). The question remains, *which* of these fairness definitions are appropriate, and in *what context*? There are two important components to answering this question: *communicating* these fairness definitions to a general audience, and *measuring their perception* of these definitions in context.

Communicating ML-related concepts is an active and growing research area. In particular, *interpretable ML* focuses on communicating the decision-making process and results of ML-based decisions to a general audience (Lipton, 2018). Many tools have been developed to make ML models more interpretable, and many demonstrably improve understanding of ML-based decisions (Ribeiro et al., 2016; Huysmans et al., 2011). These models often rely on concepts from probability and statistics—teaching these concepts has long been an active area of research. Batanero et al. (2016) provide an overview of teaching probability and how students learn probability; our surveys use their method of communicating probability, which relies on proportions. We draw on several other concepts from this literature for our study design; for example avoiding numerical and statistical representations (Gigerenzer & Edwards, 2003; Gigerenzer et al., 2007), which can be confusing to a general audience.

Instead we provide relatable examples, accompanied by examples and graphics (Hogarth & Soyer, 2015).

Effectively communicating ML concepts is necessary to achieve our second goal of understanding peoples’ perceptions of these concepts. One particularly active research area focuses on how people perceive bias in algorithmic systems. For example, Woodruff et al. (2018) investigated perceptions of algorithmic bias among marginalized populations, using a focus group-style workshop; Grgic-Hlaca et al. (2018) studies the underlying factors causing perceptions of bias, highlighting the importance of selecting appropriate features in algorithmic decision-making; Plane et al. (2017) look at perceptions of discrimination of online advertising. A related body of work studied how people perceive algorithmic decision-makers. Lee (2018) studies perceptions of fairness, trust, and emotional response of algorithmic decision-makers — as compared to human decision-makers. Similar work studies perception of fairness in the context of splitting goods or tasks (Lee & Baykal, 2017; Lee et al., 2019). Binns et al. (2018) studies how different explanation styles impact perceptions of algorithmic decision-makers.

This substantial body of prior research provided inspiration and guidance for our work. Prior work has studied both the effective communication of, and perceptions of, ML-related concepts. We hypothesize that these concepts are in fact related; to that end, we design experiments to simultaneously study peoples’ *comprehension* of and *perceptions* of common ML fairness definitions.

## 3. Methods

To study perceptions of ML fairness, we conducted two online surveys where participants were presented with a hypothetical decision-making scenario. The participants were then presented with a “rule” for enforcing fairness. We then asked each participant several questions on their comprehension and perceptions of this fairness rule. We first conducted Study-1 to validate our methodology; we then conducted the larger and broader Study-2 to address our main research questions. Both studies were approved by our organization’s standard ethical review process.

### 3.1. Study-1

In Study-1 we tested three different decision-making scenarios based on real-world decision problems: hiring, giving employee awards, and judging a student art project. However, we observed no difference in participant responses between these scenarios; for this reason we discuss only the the *hiring* decision scenario, which was also the subject of Study-2. Please see Appendix C for a description of these scenarios and survey results. In Study-1, we chose (what we believe is) the simplest definition of ML fairness—

demographic parity. In short, this rule requires that the fraction of one group who receives a *positive* outcome (i.e., an award or job offer) is equal for both groups.

### 3.1.1. SURVEY DESIGN

Here we provide a high-level discussion of the survey design; the full text of each survey can be found in Appendix C. The participant is first presented with a consent form (see Appendix D). If consent is obtained, the participant sees a short paragraph explaining the decision-making scenario. To make demographic parity accessible to a non-technical audience, and to avoid bias related to algorithmic decision-making, we frame this notion of fairness as a *rule* that the decision-maker must follow to be fair. In the hiring scenario, we framed this decision rule as follows: *The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, the fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.*

We then ask two questions concerning participant evaluation of the scenario, nine comprehension questions about the fairness rule, two self-report questions on participant understanding and use of the rule, and four free response questions on comprehension and sentiment. For example, one comprehension question is: *Is the following statement TRUE OR FALSE: This hiring rule always allows the hiring manager to send offers exclusively to the most qualified applicants.* Finally, we collect demographic information (age, gender, race/ethnicity, education level, and expertise in a number of relevant fields).

We conducted in-person cognitive interviews (Harrell & Bradley, 2009) to pilot our survey, leading to several improvements in the question design. Most notably, because some cognitive interview participants appeared to use their own personal notions of fairness rather than our provided rule, we added questions to assess this compliance issue.

### 3.1.2. RECRUITMENT AND PARTICIPANTS

We recruited participants using the online service Cint (Cint), which allowed us to loosely approximate the 2017 U.S. Census distributions (Bureau, 2017) for ethnicity and education level, allowing for broad representation. We required that participants be 18 years of age or older, and fluent in English. Participants were compensated using Cint's rewards system; according to a Cint representative: "[Participants] can choose to receive their rewards in cash sent to their bank accounts (e.g. via PayPal), online shopping opportunities with one of multiple online merchants, or donations to a charity."

In total 147 participants were included in the Study-1 analysis, including 75 men (51.0%), 71 women (48.3%), and 1

(0.7%) preferring not to answer. The average age was 46 years ( $SD = 16$ ). Ethnicity and educational attainment are summarized in Table 1. On average, participants completed the survey in 14 minutes.

Table 1 summarizes the ethnicity and education level of participants in both Study-1 and Study-2.

	Census	Percent of Sample Study-1	Percent of Sample Study-2
<b>Ethnicity</b>			
AI or AN	0.7	0.7	0.9
Asian or NH or PI	5.7	1.4	2.3
Black or AA	12.3	10.2	15.8
Hispanic or Latinx	18.1	12.2	7.7
Other	2.6	2.7	1.4
White	60.6	72.8	71.9
<b>Education Level</b>			
Less than HS	12.1	6.1	6.9
HS or equivalent	27.7	29.9	24.9
Some post-secondary	30.8	30.6	24.9
Bachelor's and above	29.4	33.3	42.7

Table 1. Participant demographics across ethnicity and education level, compared to the 2017 U.S. Census. AI = American Indian, AN = Alaska Native, NH = Native Hawaiian, PI = Pacific Islander, AA = African American. Note that in Study-2, two participants did not report their education level.

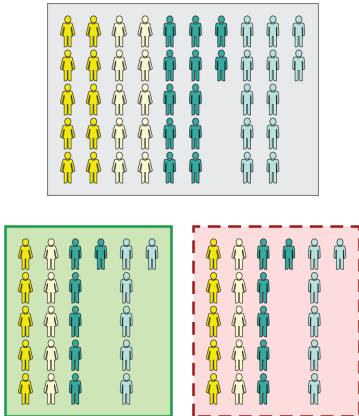
## 3.2. Study-2

Study-2 follows a very similar structure to Study-1 with a few changes. First, we decided to use only the hiring (HR) decision scenario (See Appendix A.5 for more in-depth discussion). Second, we expanded to three definitions of fairness: *demographic parity* (DP), *equal opportunity* (EP), and *equalized odds* (EO) (Hardt et al., 2016). Within EP, we tested both False Negative Rate (FNR) and False Positive Rate (FPR), resulting in a total of four conditions.

### 3.2.1. SURVEY DESIGN

Here we provide a high-level discussion of the differences between Study-2 and Study-1; the full text of each survey can be found in Appendix C. We used a between-subjects design with random assignment among the four conditions (DP, FNR, FPR, EO). Again, we frame each notion of fairness as a *hiring rule* that the decision-maker must follow to be fair. For example, in FPR we define the award rule as follows: *The fraction of unqualified male candidates who receive job offers should equal the fraction of unqualified female candidates who receive job offers.*

For this version, we added graphical examples to further clarify our explanations (see Fig. 1 for an example). We used the all the same questions as in Study-1 but added two additional Likert scale questions assessing whether the participant sentiment: one asked whether they liked the rule,



**Figure 1.** A graphical example to describe a fair hiring outcome for EO. Yellow people represent females while green people represent males. The darker colors represent qualified individuals while the lighter colors represent unqualified individuals. The gray box represents the original pool of applicants. The green box represent individuals that received job offers while the red box with a dashed border represents individuals that *not* receive job offers.

and the other asked whether they agreed with the rule. One free response question (asking how participants personally would go about the hiring process to ensure it was fair), which did not consistently provide useful responses in Study-1, was removed from the Study-2 survey in an effort to keep the expected completion time similar.

### 3.2.2. RECRUITMENT AND PARTICIPANTS

We again used the Cint service to recruit participants. Because our initial sample (intended to target education, ethnicity, gender and age distributions approximating the U.S. census) skewed more highly educated than we had hoped, we added a second round one week later primarily targeting participants without bachelor's degrees. Hereafter, we report on both samples together.

In total 349 participants were included in the Study-2 analysis, including 142 men (40.7%), 203 women (58.2%), 1 other (0.3%), and 3 (0.9%) preferring not to answer. The average age was 45 years ( $SD = 15$ ). Ethnicity and educational attainment are summarized in Table 1. On average, participants completed the survey in 16 minutes.

### 3.3. Data Analysis

Free response questions were qualitatively coded for statistical testing. In Study-1, one question was coded by a single researcher for simple correctness (see Appendix A.1), and the other was independently coded by three researchers (resolved to 100%) to capture sentiment information (see Appendix A.3). In Study-2, both questions were independently coded by 2-3 researchers (resolved to 100%). Participants who provided nonsensical answers, answers not in English,

or other non-responsive answers to free response questions were excluded from all analysis.

The following methods were used for all statistical analyses unless otherwise specified. Correlations with nonparametric ordinal data were assessed using Spearman's rho. Omnibus comparisons on nonparametric ordinal data were performed with a Kruskal-Wallis (K-W) test, and relevant post-hoc comparisons with Mann-Whitney U (M-WU) tests. Post-hoc  $p$ -values were adjusted for multiple comparisons using Bonferroni correction.  $\chi^2$  tests were used for comparisons of nominal data. Boxplots show median and first and third quartiles; whiskers extend to  $1.5 * IQR$  (interquartile range), with outliers indicated by points.

### 3.4. Limitations

As with all surveys, our study has certain limitations. We recruited a demographically broad population, but web panels are generally more tech-savvy than the broader population (Redmiles et al., 2019). We consider this acceptable for a first effort. Some participants may be satisficing rather than answering carefully. We mitigate this by disqualifying participants with off-topic or non-responsive free-text responses. Further, this limitation can be expected to be consistent across conditions, enabling reasonable comparison. Finally, better or clearer explanations of the fairness definitions we explored are certainly possible; we believe our explanations were sufficient to allow us to investigate our research questions, especially because they were designed to be consistent across conditions.

## 4. Results

In this section we first discuss the preliminary findings from Study-1 (see §4.1). These findings were used as hypotheses for further exploration and testing in Study-2; we discuss those results second (see §4.2).

### 4.1. Study-1

We analyze survey responses for Study-1 and make several observations. We first validate our comprehension score as a measure of participant understanding; we then generate hypotheses for further exploration in Study-2.

#### 4.1.1. OUR SURVEY EFFECTIVELY CAPTURES RULE COMPREHENSION

We find that we can measure comprehension of the fairness rule. The comprehension score was calculated as the total correct responses out of a possible 9. All questions were weighted equally. The relevant questions included 2 multiple choice, 4 true/false, and 3 yes/no questions. The average score was 6.2 ( $SD=2.3$ ).

We validate our comprehension score using two methods: internal validity testing, and correlation against two self-report and one free response question included in our survey (see Appendix A.1 for further details).

**Internal Validity** Cronbach's  $\alpha$  and item-total correlation were used to assess internal validity of the comprehension score. Both measures met established thresholds (Nunnally, 1978; Everitt & Skrondal, 2010): Cronbach's  $\alpha = 0.71$ , and item-total correlation for 8 of the 9 items (all but Q5)  $> 0.3$ .

**Question Correlation** We find that self-reported rule understanding and use are reflected in comprehension score. First, we compared comprehension score to self-reported rule understanding (Q13): "I am confident I know how to apply the award rule described above," rated on a five-point Likert scale from strongly agree (1) to strongly disagree (5). The median response was "agree" (Q1 = 1, Q3 = 3). Higher comprehension scores tended to be associated with greater confidence in understanding (Spearman's  $\rho = 0.39$ ,  $p < 0.001$ ), supporting the notion that comprehension score is a valid measure of rule comprehension.

Next, we compared comprehension score to a self-report question about the participant's use of the rule (Q14), with the following options: (a) "I applied the provided award rule only," (b) "I used my own ideas of what the correct award decision should be rather than the provided award rule," or (c) "I used a combination of the provided award rule and my own ideas of what the correct award decision should be." We find that participants who claimed to use only the rule scored significantly higher (mean 7.09) than those who used their own notions (4.68) or a combination (4.90) (post-hoc M-WU,  $p < 0.001$  for both tests; corrected  $\alpha = 0.05/3 = 0.017$ ). This further corroborates our comprehension score.

Finally, we asked participants to explain the rule in their own words (Q12). Each response was then qualitatively coded as one of five categories – **Correct**: describes rule correctly; **Partially correct**: description has some errors or is somewhat vague; **Neither**: vague description of purpose of the rule rather than how it works, or pure opinion; **Incorrect**: incorrect or irrelevant; and **None**: no answer, or expresses confusion. Participants whose responses were either correct (mean comprehension score = 7.71) or partially correct (7.03) performed significantly better on our survey than those responding with neither (5.13) or incorrect (4.24) (post-hoc M-WU,  $p < 0.001$  for these four comparisons, corrected  $\alpha = 0.005$ ). These findings further validate our comprehension score. Additional details of these results and the associated statistical tests can be found in Appendix A.1.

#### 4.1.2. HYPOTHESES GENERATED

We analyzed the data from Study-1 in an exploratory fashion intended to generate hypotheses that could be tested

in Study-2. We highlight here three key hypotheses that emerged from the data.

**Education Influences Comprehension** We used poisson regression models to explore whether various demographic factors were associated with differences in comprehension. We found that a model including education as a regressor had greater explanatory power than a model without (see Appendix A.2 for further details).

**Disagreement with the Rule is Associated with Higher Comprehension Scores** We asked participants for their opinion on the presented rule in a free response question (Q15). These responses were qualitatively coded to capture participant sentiment toward the rule in one of five categories – **Agree**: generally positive sentiment towards rule; **Depends**: describes both pros and cons of the given rule; **Disagree**: generally negative sentiment towards rule; **Not understood**: expresses confusion about rule; **None**: no answer, or lacks opinion on appropriateness of the rule. Participants who expressed disagreement with the rule performed better (mean comprehension score = 7.02) than those who expressed agreement (5.50), did not understand the rule (4.44), or provided no response (5.09) to the question (post-hoc M-WU,  $p < 0.005$  for these three comparisons; corrected  $\alpha = 0.05/10 = 0.005$ ). Appendix A.3 provides further details.

**Non-Compliance is Associated with Lack of Understanding** We were interested in understanding why some participants failed to adhere to the rule, as measured by their self-report of rule usage in Q14. We labeled those who responded with either having used their own personal notions of fairness ( $n = 29$ ) or some combination of their personal notions and the rule ( $n = 28$ ) as "non-compliant" (NC), with the remaining  $n = 89$  labeled as "compliant" (C). One participant who did not provide a response was excluded from this analysis, conducted using  $\chi^2$  tests.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (see Fig. 13). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 (see Fig. 14). This fits with the overall strong relationship we observed among comprehension scores, self-reported understanding, ability to explain the rule, and compliance.

Further, negative participant sentiment towards the rule (Q15) also appears to be associated with greater compliance (see Fig. 15). Thus, non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it. Refer to Appendix A.4 for further details.

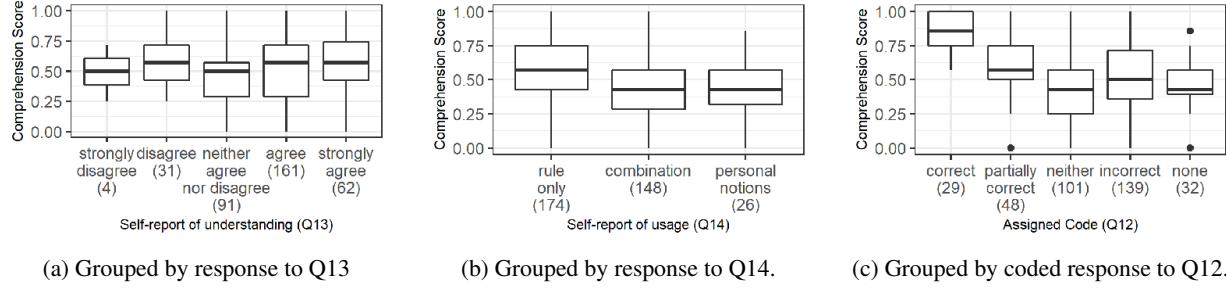


Figure 2. Comprehension scores grouped by questions. In (a), self-reported understanding of the rule was not related to comprehension score. X-axis is reversed for figure and correlation test. In (b), rule compliance (leftmost on the x-axis) was associated with higher comprehension scores. One participant who did not provide a response was excluded from this figure and the relevant analysis. Finally, in (c), participants who provided either correct or partially correct responses tended to perform better.

## 4.2. Study-2

We first confirm the validity of our comprehension score, then compare comprehension across definitions and examine the hypotheses generated in Study-1.

### 4.2.1. SCORE VALIDATION

We validated our metric using the same approach used in Study-1, i.e., assessing both internal validity and correlation with self-report and free-response questions. We report the results of this assessment here.

**Internal Validity** We again used Cronbach’s  $\alpha$  and item-total correlation to assess internal validity of the comprehension score. An initial assessment using all 349 responses yielded Cronbach’s  $\alpha = 0.38$ , and item-total correlation  $> 0.3$  for only four of the nine comprehension questions. Since both measures performed below established thresholds (Nunnally, 1978; Everitt & Skrondal, 2010), we investigated further and repeated these measurements individually for each fairness-definition condition (DP, FNR, FPR, EO). This procedure showed stark differences in Cronbach’s  $\alpha$  based on definition: DP = 0.64, FNR = 0.39, FPR = 0.49, EO = 0.62. Item-total correlations followed a similar pattern: best in DP, worst in FNR. Based on these differences, we iteratively removed problematic questions from the score on a per-definition basis until all remaining questions achieved an item-total correlation of  $> 0.3$  (Everitt & Skrondal, 2010). By removing poorly performing questions, we increase our confidence that the measured comprehension scores are meaningful for further analysis. Table 2 specifies which questions were retained for analysis in each definition.

Because questions were dropped on a per-definition basis, the range of the resulting scores varied from 4-7 depending on the definition, rather than being a uniform 9. We normalized this treating comprehension score as a percentage of the maximum for each condition rather than a raw score. We report this *adjusted score* in the remainder of §4.2.

	Questions									
	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	
DP	X	X				X	X	X	X	
FNR	X	X	X		X					
FPR	X	X	X	X		X		X	X	
EO	X		X		X	X	X	X	X	

Table 2. Questions that were used for downstream analysis after iterative removal of questions with poor item-total correlation.

**Question Correlation** As in Study-1, we compare comprehension scores with responses to self-report and free response questions included in our survey.

First, we compared comprehension score to self-reported rule understanding (Q13), as described in §4.1.1. The median response was “agree” (Q1 = 2, Q3 = 3). We assess the correlation between these responses and comprehension score using Spearman’s rho (appropriate for ordinal data). Unlike in Study-1, there was no relationship between self-reported understanding and comprehension score (Fig. 2a).

Next, we compared comprehension score to a self-report question about the participant’s use of the rule (Q14), as described in §4.1.1. A K-W test revealed a relationship between self-reported rule usage and comprehension score ( $p < 0.001$ ). We find that participants who claimed to use only the rule tended to score higher (mean comprehension score = 0.60) than those who used their own notions (0.47) or a combination (0.45) thereof (post-hoc M-WU,  $p < 0.01$  for both tests; corrected  $\alpha = 0.05/3 = 0.017$ ). This suggests that participants are answering at least somewhat honestly: when they try to apply the rule, comprehension scores improve (see Fig. 2b).

Finally, we asked participants to explain the rule in their own words (Q12). Each response was then qualitatively coded as one of five categories, as described in S4.1.1. These results can be seen in Fig. 2c. A K-W test revealed a relationship between comprehension score and coded responses to Q12 ( $p < 0.001$ ). Correct (mean comprehension score = 0.86) re-

Covariate	Est.	95% CI	p
<i>Education</i>			
HS	0.02	[-0.08, 0.12]	0.720
Post-secondary, no BS	0.09	[-0.01, 0.19]	0.091
Bachelor's and above	0.17	[0.08, 0.27]	< 0.001
<i>Definition</i>			
EO	-0.04	[-0.11, 0.02]	0.218
FPR	-0.05	[-0.11, 0.02]	0.138
FNR	-0.14	[-0.20, -0.07]	< 0.001

Table 3. Regression table for the best fit model, with two covariates: education (baseline: no HS) and definition (baseline: DP). Est. = estimate, CI = confidence interval.

sponses were associated with higher comprehension scores than partially correct (0.60), neither (0.44), incorrect (0.52), and none (0.46) responses ( $p < 0.001$  for all); partially correct responses were also associated with higher comprehension scores than neither and none responses ( $p < 0.001$  for both). No other differences were found (post-hoc MWU; corrected  $\alpha = 0.05/10 = 0.005$ ). These findings support our claim that our comprehension score is a valid measure of fairness-rule comprehension.

#### 4.2.2. EDUCATION AND DEFINITION ARE RELATED TO COMPREHENSION SCORE

One hypothesis generated by Study-1 was that comprehension score is positively correlated with education level. We investigated this hypothesis using linear regression models.

Eleven models were tested, regressing different combinations of demographics (ethnicity, gender, education, and age) and condition (fairness definition). Models were compared using Akaike information criterion (AIC), a standard method of evaluating model quality and performing model selection (Akaike, 1974). Comparison by AIC revealed that the model using just education (edu) and fairness definition (def) as regressors was the model of best fit. In this model, having a Bachelor's degree or above resulted in a score increase of 0.17, and the FNR condition caused a score decrease of -0.14 ( $p < 0.001$  for both; corrected  $\alpha = 0.05/11 = 0.0045$ ). A regression table of the best fit model is below, in Table 3.

AIC results of each of the eleven models, along with the relevant regressors, can be seen in Table 4 in Appendix B.1. Comprehension score as a function of education and fairness definition can be seen in Figs. 3 and 4.

#### 4.2.3. GREATER NEGATIVE SENTIMENT TOWARD THE RULE IS ASSOCIATED WITH HIGHER COMPREHENSION SCORES

In Study-1, we found a relationship between participant sentiment towards the rule and comprehension score. To better interrogate this phenomenon, in Study-2 we added

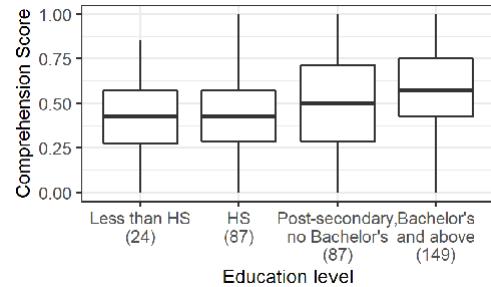


Figure 3. Comprehension score grouped by education level. Higher education was associated with higher comprehension scores. Note that two participants who did not report their education level were removed from this figure and the relevant analysis.

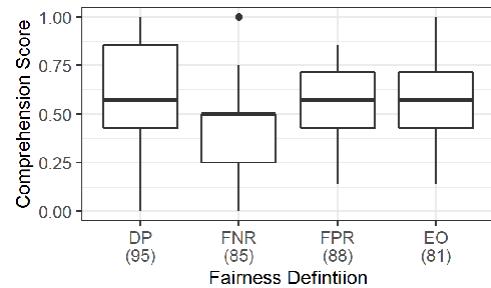


Figure 4. Comprehension score grouped by fairness definition. The FNR condition was associated with lower comprehension score.

two more questions to the survey to directly address the issue of sentiment, rather than relying on a free response question. One (Q15) asks, “To what extent do you agree with the following statement: I like the hiring rule?”, and is evaluated on a five-point Likert scale from “strongly agree” (1) to “strongly disagree” (5). The other (Q16) asks, “To what extent do you agree with the following statement: I agree with the hiring rule?”, and is also evaluated on a five-point Likert scale from “strongly agree” (1) to “strongly disagree” (5).

Using Spearman’s rho, we assessed the correlation between responses to these two questions and comprehension score. A minor correlation was found between liking the rule and comprehension score, in that those who disliked the rule were more likely to have higher comprehension scores ( $\rho = -0.15, p < 0.01$ ; see Fig. 5). No correlation was found between agreeing with the rule and comprehension score (see Fig. 6).

#### 4.2.4. NON-COMPLIANCE IS ASSOCIATED WITH LACK OF UNDERSTANDING

A final hypothesis generated in Study-1 involves non-compliance: i.e., why do participants who report *not* using the rule to answer the comprehension questions behave this way? In Study-1, we found that this was due to the fact that non-compliant participants were less able to *understand* the rule, rather than because they did not *like* it. We also

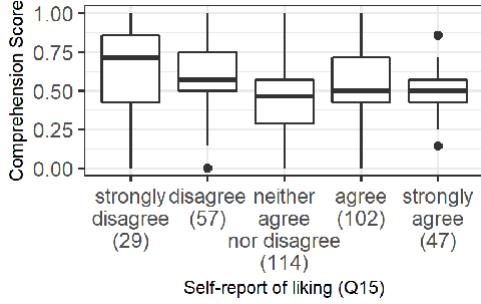


Figure 5. Comprehension score grouped by response to Q15. Disliked of the rule was associated with higher comprehension scores. X-axis is reversed for figure and correlation test.

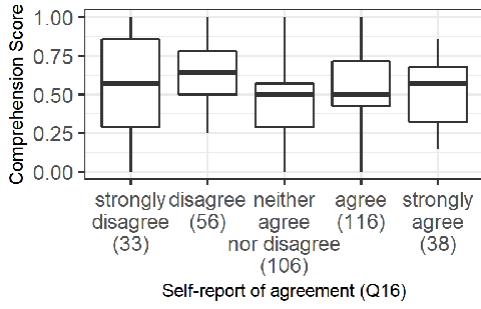


Figure 6. Comprehension score grouped by response to Q16. Rule agreement was not correlated with comprehension score. X-axis is reversed for figure and correlation test.

observed this in our results form Study-2: compliant participants exhibited higher self-reported understanding of the rule ( $p < 0.001$ , Fig. 17), were more likely to correctly explain the rule ( $p < 0.001$ , Fig. 18), and were significantly more likely to dislike the rule ( $p < 0.05$ , Fig. 19). Refer to Appendix B.2 for more details. As with comprehension score, we observed no relationship between compliance and agreement with the rule (Fig. 20).

## 5. Discussion

Bias in machine learning is a growing threat to justice; to date, ML bias has been documented in both commercial and government applications, in sectors such as medicine, criminal justice, and employment. In response, ML researchers have proposed various notions of *fairness* to correct these biases. Most ML fairness definitions are purely mathematical, and require some knowledge of machine learning. While they are intended to benefit the general public, it is unclear whether the general public agrees with — or even understands — these notions of ML fairness.

We take an initial step to bridge this gap by asking *do people understand the notions of fairness put forth by ML researchers?* To answer this question we develop a short questionnaire to assess understanding of three particular notions of ML fairness (demographic parity, equal opportunity, and equalized odds). We find that our comprehension score

(with some adjustments for each definition) appears to be a consistent and reliable indicator of understanding the fairness metrics. The comprehension score demonstrated in this work lays a foundation for many future studies exploring other fairness definitions.

We do find, however, that comprehension is lower for equal opportunity, false negative rate than other definitions. In general, comprehension scores for equal opportunity (both FNR and FPR) were less internally consistent than other fairness rules, suggesting participant responses were also more “noisy” for equal opportunity. This is somewhat intuitive: equal opportunity is difficult to understand, as it only involves one type of error (FNR or FPR) rather than both. Furthermore, FNR participants had the lowest comprehension scores *and* the lowest consistency of all conditions. We believe this finding also matches intuition: FNR is a strange notion in the context of hiring, as it concerns only those who were *not* hired or offered jobs. Indeed, in free-response questions several participants mentioned that they do not understand why qualified candidates are *not* hired. We believe many participants fixated on this strange setting, impacting their comprehension scores. This finding is potentially problematic, as equal opportunity definitions are increasingly used in practice. Indeed, major fairness tools such as Google What-If tool (Wexler et al., 2019) and the IBM AI Fairness 360 (Bellamy et al., 2019) specifically focus on equal opportunity. Further work should be put into making descriptions of nuanced fairness metrics more accessible.

Our analysis also identified other issues that should be considered when thinking about mathematical notions of fairness. First, we find that education is a strong predictor of comprehension. This is especially troubling, as the negative impacts of biased ML are expected to disproportionately impact the most marginalized (Barocas & Selbst, 2016) and displace employment opportunities for those with the least education (Frey & Osborne, 2017). Lack of understanding may hamper these groups’ ability to effectively advocate for themselves. Designing more accessible explanations of fairness should be a top research priority.

Second, we find that those with the weakest comprehension of fairness metrics also express the least negative sentiment toward them. When fairness is a concern, there are always trade-offs — between accuracy and equity, or between different stakeholders, and so on. Balancing these trade-offs is an uncomfortable dilemma often lacking an objectively correct solution. It is possible that those who comprehend this dilemma *also* recognize the precarious trade-off struck by any mathematical definition of fairness, and are therefore dissatisfied with it. From another perspective, this finding is more insidious. If those with the weakest understanding of AI bias are also least likely to protest, then major problems in algorithmic fairness may remain uncorrected.

## References

- Akaike, H. A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, volume 19, pp. 716723, 1974. doi: 10.1109/TAC.1974.1100705.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica, May*, 23:2016, 2016.
- Baracas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., and Sánchez, E. Research on teaching and learning probability. In *Research on teaching and learning probability*, pp. 1–33. Springer, Cham, 2016.
- Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.
- Binns, R. Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, 81:1–11, 2017.
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. ‘It’s reducing a human being to a percentage’: Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 377. ACM, 2018.
- Bureau, U. C. 2017 us census demographics, 2017. URL <https://data.census.gov/cedsci>.
- Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- Chouldechova, A. and Roth, A. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.
- Cint. Cint. URL <https://www.cint.com/>.
- Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.
- Everitt, B. and Skrondal, A. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 4th edition, 2010.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.
- Frey, C. B. and Osborne, M. A. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.
- Gigerenzer, G. and Edwards, A. Simple tools for understanding risks: from innumeracy to insight. *Bmj*, 327(7417):741–744, 2003.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96, 2007.
- Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pp. 903–912. International World Wide Web Conferences Steering Committee, 2018.
- Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS*, pp. 3315–3323, 2016.
- Harrell, M. C. and Bradley, M. A. Data collection methods. semi-structured interviews and focus groups. Technical report, Rand National Defense Research Inst santa monica ca, 2009.
- Hogarth, R. M. and Soyer, E. Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition*, 4(3):221–228, 2015.
- Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis. Support Syst.*, 51(1):141–154, April 2011. ISSN 0167-9236. doi: 10.1016/j.dss.2010.12.003. URL <http://dx.doi.org/10.1016/j.dss.2010.12.003>.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.
- Lee, M. K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018.

- Lee, M. K. and Baykal, S. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pp. 1035–1048, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998230. URL <http://doi.acm.org/10.1145/2998181.2998230>.
- Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. In *Proc. ACM Hum.-Comput. Interact.*, 3, CSCW, pp. Article 182, New York, NY, USA, 2019. ACM. URL <https://doi.org/10.1145/3359284>.
- Lipton, Z. C. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.
- Nunnally, J. *Psychometric Theory*. McGraw-Hill, 2nd edition, 1978.
- Plane, A. C., Redmiles, E. M., Mazurek, M. L., and Tschantz, M. C. Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)*, pp. 935–951, 2017.
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.
- Rawls, J. *A Theory of Justice*. Harvard University Press, 1971.
- Redmiles, E. M., Kross, S., and Mazurek, M. L. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 1326–1343. IEEE, 2019.
- Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.
- Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.
- Woodruff, A., Fox, S. E., Rousso-Schindler, S., and Warshaw, J. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 656. ACM, 2018.

## A. Study-1: Detailed Results

### A.1. Our Survey Effectively Captures Rule Comprehension

We find that our survey is internally consistent, and effectively measures participant comprehension of demographic parity. The former we evaluated using Cronbach's  $\alpha$  and item-total correlation (discussed in §4.1.1), and the latter using two self-report measures and one free response question. See Fig. 7 for participant performance per question.

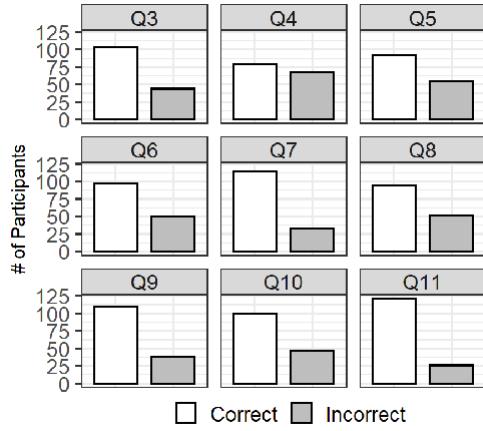


Figure 7. Number of participants answering each question correctly. Each panel contains all 147 participants.

#### A.1.1. SELF-REPORTED RULE UNDERSTANDING AND USE ARE REFLECTED IN COMPREHENSION SCORE

First, we compared comprehension score to self-reported rule understanding (Q13). Higher comprehension scores were associated with greater confidence in understanding (Spearman's rho), suggesting that participants were accurately assessing their ability to apply the rule (see Fig. 8).

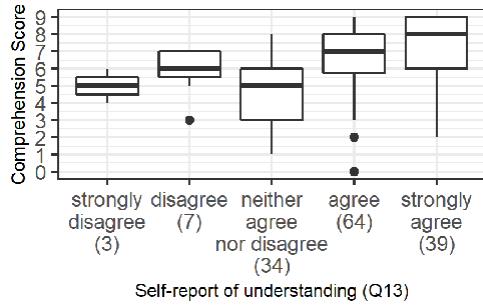


Figure 8. Comprehension score grouped by response to Q13. Self-reported understanding of the rule was associated with higher comprehension scores. X-axis is reversed for figure and correlation test.

Next, we compared comprehension score to a self-report question about the participant's use of the rule (Q14). Participants who claimed to use only the rule tended to score

higher than those who used their own notions of fairness or a combination thereof (K-W test, and post-hoc M-WU), suggesting that participants are answering somewhat honestly: when they try to apply the rule, comprehension scores improve (see Fig. 9).

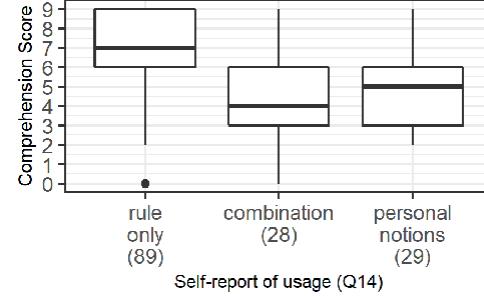


Figure 9. Comprehension score grouped by response to Q14. Rule compliance (leftmost on the x-axis) was associated with higher comprehension scores. One participant who did not provide a response was excluded from the figure and relevant analysis.

#### A.1.2. PARTICIPANTS WITH HIGHER COMPREHENSION SCORES ARE BETTER ABLE TO EXPLAIN THE RULE

To further validate our comprehension score, we asked participants to explain the rule in their own words (Q12). Responses were qualitatively coded as one of five categories: **correct**, **partially correct**, **neither**, **incorrect**, or **none** (as discussed in §4.1.1). The results of this coding can be seen in Fig. 10. Participants providing correct explanations of the rule attained higher comprehension scores (k-W test, and post-hoc M-WU), further corroborating our claim that our comprehension score is a valid measure of fairness rule comprehension.

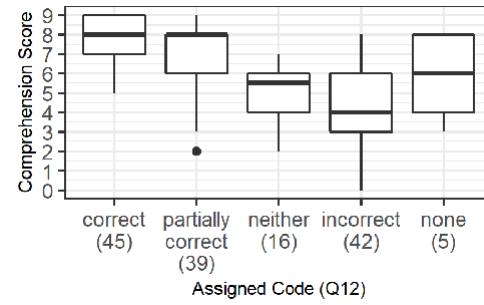


Figure 10. Comprehension score grouped by code assigned to Q12 response. Participants who provided either correct or partially correct responses tended to perform better.

## A.2. Education Influences Comprehension

During the cognitive interview phase, we observed a possible trend of comprehension scores being lower for older participants and those with less educational attainment. If

true, this would suggest that fairness explanations should be carefully validated to ensure they can be used with diverse populations. We investigated this hypothesis, in an exploratory fashion, using poisson regression models.

Three models were tested. The first regressed score against all four demographic categories as predictors (gender, age, ethnicity, and education), the second omitted education, and the third tested only education. Models were compared using Akaike information criterion (AIC), a standard method of evaluating model quality and performing model selection (Akaike, 1974). Comparison by AIC revealed that model 1 (all four categories) was a better predictor for comprehension score than models 2 or 3 (AIC = 643.3, 651.2, and 660.5, respectively; difference = 0.0, 7.9, and 17.1). In model 1, only education showed correlation with comprehension score (effect size = 1.40,  $p < 0.05$ ). Further work is needed to confirm this exploratory result.

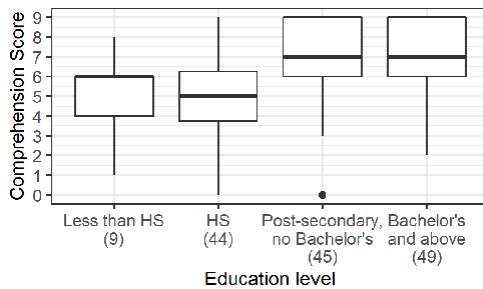


Figure 11. Comprehension score grouped by education level. Higher education level was associated with higher comprehension scores.

### A.3. Disagreement with the Rule is Associated with Higher Comprehension Scores

Participants were asked for their opinion on the presented rule in another free response question (Q15). These responses were then qualitatively coded to capture participant sentiment towards the rule as one of five categories: **agree**, **depends**, **disagree**, **not understand**, or **none** (as discussed in §4.1.2).

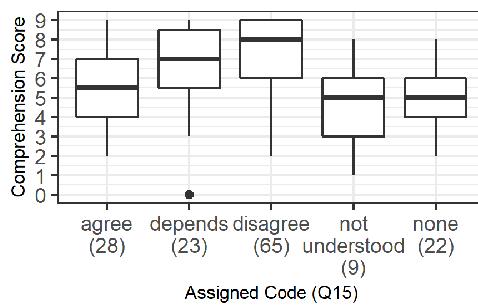


Figure 12. Comprehension score grouped by code assigned to Q15 response. Participants who exhibited negative sentiment toward the rule responses tended to perform better.

This question was added based on the cognitive interviews (see Appendix E.1), where perception seemed to influence compliance. The results of coding Q15 can be seen in Fig. 12. Participants who expressed disagreement with the rule performed better than those who expressed agreement, did not understand the rule, or provided no response to the question (K-W test, post-hoc M-WU). Note that this result should not be interpreted as an overall finding on the appropriateness of demographic parity. Instead we anticipate the perceptions of appropriateness of any fairness definition will be highly context-dependent.

### A.4. Non-Compliance is Associated with Lack of Understanding

We were interested in understanding why some participants failed to adhere to the rule, as measured by their self-report of rule usage in Q14. After labeling participants as either “non-compliant” (NC,  $n = 57$ ) or “compliant” (C,  $n = 89$ ), we conducted a series of  $\chi^2$  tests to investigate this phenomenon.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (see Fig. 13). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 (see Fig. 14). Further, negative participant sentiment towards the rule (Q15) also appears to be associated with greater compliance (see Fig. 15). Thus, non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it.

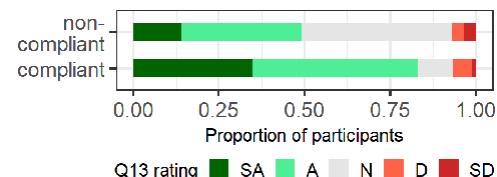


Figure 13. Self-report of understanding (Q13) split by compliance (Q14). NC participants tend to report less confidence in their ability to apply the rule. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

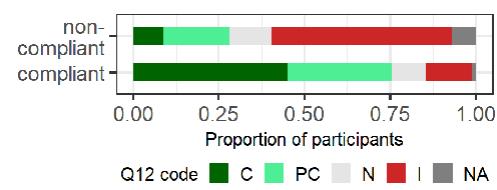


Figure 14. Correctness of rule explanation (Q12) split by compliance (Q14). NC participants tend to be less able to explain the presented rule in their own words. C = correct, PC = partially correct, N = neither, I = incorrect, NA = none.

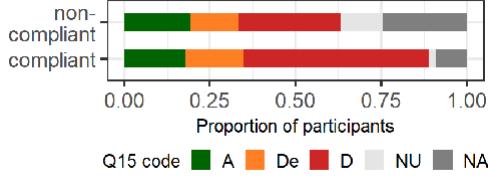


Figure 15. Participant agreement with rule (Q15) split by compliance (Q14). NC participants tend to harbor less negative sentiment towards the rule. A = agree, De = depends, D = disagree, NU = not understood, NA = none.

### A.5. Decision Scenarios

For Study-1 we designed three decision-making scenarios to test whether the perceived importance or realism of a particular scenario influenced comprehension score. They are as follows:

- **Art Project (AP):** distributing awards for art projects to primary school students,
- **Employee Awards (EA):** distributing employee awards at a sales company, and
- **Hiring (HR):** distributing job offers to applicants.

In each scenario the students/employees/applicants are partitioned into two groups (parents’ occupation for the first scenario, and binary gender for the other two scenarios). We use a between-subjects design: participants are randomly partitioned into three conditions, one for each scenario (AP, EA, or HR). For each condition we define the *fairness rule* in the context of the decision-making scenario (see Appendix C for the full surveys).

Next we describe our main conclusion related to the different decision-making scenarios in Study-1: the scenario does not influence comprehension score.

#### A.5.1. SCENARIO DOES NOT INFLUENCE COMPREHENSION SCORES (RQ4)

We were concerned that less important and/or realistic scenarios would cause participants to take the survey less seriously, and therefore perform more poorly. To test this, participants were randomly assigned to a scenario, resulting in the following distribution: AP = 41, EA = 49, HR = 57.

A K-W test revealed no differences between scenarios in terms of comprehension score (mean comprehension scores: AP = 6.0, EA = 6.74, HR = 5.86). However, differences did exist between scenarios in terms of importance (assessed in Q2), measured in hours of effort deemed necessary to make the relevant decision (K-W,  $p < 0.001$ ). Post-hoc M-WU revealed that participants believed making a decision in the AP scenario merited fewer hours of effort (mean = 3.15hrs) than in the EA (13.52hrs,  $p < 0.001$ ) or HR (15.23hrs,

$p < 0.001$ ) scenarios (corrected  $\alpha = 0.05/3 = 0.017$ ). See Fig. 16 for distributions of responses.

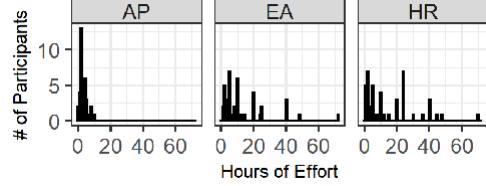


Figure 16. Importance of a scenario by proxy of hours of effort necessary to make a decision in each scenario. AP merited less hours of effort than both EA and HR.

Of note, it is possible that perceived realism, assessed in Q1 on a five-point Likert scale, was also influenced by scenario (K-W,  $p = 0.051$ ), but we may need larger sample sizes to confirm this. Regardless, while the nature of a scenario does influence participant perception in terms of importance and (possibly) realism, it does not appear to influence comprehension (at least for the scenarios we chose). For this reason, we chose to test a single scenario (HR) in Study-2.

## B. Study-2: Detailed Results

### B.1. Model Selection

In §4.2.2 we assessed eleven linear regression models for predicting comprehension scores. The best fit model, determined by model selection via AIC, included only education (edu) and fairness definition (def) as regressors. The results of model selection are below in Table 4.

Model regressors	AIC	dAIC
edu + def	-51.0	0.0
edu	-39.1	12.0
gender + edu	-36.2	14.9
gender + age + eth + edu + def	-33.8	17.2
age + edu	-30.5	20.5
gender + age + edu	-27.6	23.4
def	-25.7	25.4
gender + age + eth + edu	-23.8	27.3
gender + age + def	-11.1	39.9
gender + age + eth + def	-8.4	42.6
gender + age + eth	1.1	52.1

Table 4. Models tested in §4.2.2, sorted by best to least fit. The first model in the table (edu + def) is the model of best fit. dAIC = difference from model with lowest AIC value.

### B.2. Non-Compliance

In §4.2.4 we sought to further investigate the findings of Study-1 with regards to compliance (Q14). To do so, we labeled those who responded (in Study-2) with either having used their own personal notions of fairness ( $n = 26$ ) or some combination of their personal notions and the rule ( $n = 148$ ) as “non-compliant” (NC), with the remaining

$n = 174$  labeled as “compliant” (C). One participant who did not provide a response was excluded from this analysis, conducted using KW and  $\chi^2$  tests.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (KW test,  $p < 0.001$ , see Fig. 17). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 ( $\chi^2$  test,  $p < 0.001$ , see Fig. 18). This fits with the overall strong relationship we observed among comprehension scores, ability to explain the rule, and compliance.

Further, greater dislike towards the rule (Q15) also appears to be associated with greater compliance (KW test,  $p < 0.05$ , see Fig. 19). However, there was no relationship between disagreement towards the rule (Q16) and compliance (see Fig. 20).

These results largely corroborate the notion that non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it.

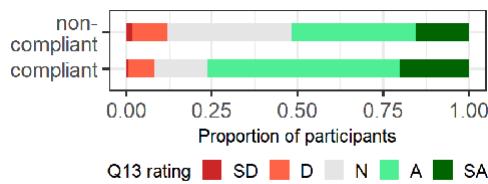


Figure 17. Self-report of understanding (Q13) split by compliance (Q14). NC participants tend to report less confidence in their ability to apply the rule. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

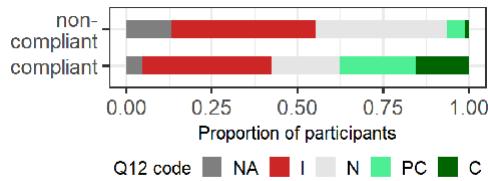


Figure 18. Correctness of rule explanation (Q12) split by compliance (Q14). NC participants tend to be less able to explain the presented rule in their own words. C = correct, PC = partially correct, N = neither, I = incorrect, NA = none.

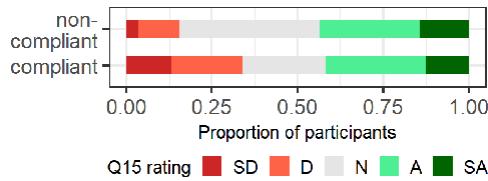


Figure 19. Participant liking for rule (Q15) split by compliance (Q14). NC participants tend to dislike the rule less than C participants. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

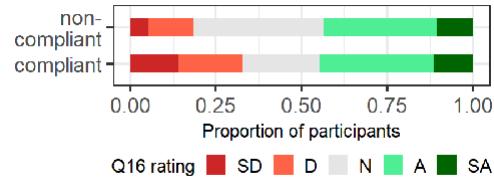


Figure 20. Participant agreement with rule (Q16) split by compliance (Q14). No differences were found between NC and C participants. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

## C. Surveys

### C.1. Study-1 Survey

Each of the surveys are split into four main sections. The first section is the consent form which can be found in Appendix D. The second section describes the scenario and asks questions about the given scenario (§C.1.1). The third section describes the fairness metric, defined as the rule, used (in this case it is demographic parity) and asks specific questions about the metric (§C.1.2). Finally the last section asks for demographic information (§C.3).

#### C.1.1. SCENARIO DESCRIPTIONS AND QUESTIONS

The following is shown to each participant:

It is very important that you read each question carefully and think about your answers. The success of our research relies on our respondents being thoughtful and taking this task seriously.

I have read the above instructions carefully.

We then introduce one of three different decision making scenarios, described below, followed by two questions. Words that vary across scenario in the questions are shown as *<art project, employee awards, hiring>*.

**Art project** A fourth grade teacher is reviewing 20 student art projects. They will award lollipops to the top 4 students who put the most effort into their projects. The teacher knows that some of the students have artists as parents, who might have helped their children with their art project. The teachers goal is to give out lollipops only based on the amount of effort that the student *themselves* put into their projects.

The teacher uses the following criteria to decide who should get a lollipop:

- Elaborateness of each project.
- Creativity of each project.

About 50% of the students have artists as parents, and 50% do not.

In the past, students with artists as parents typically put more effort into their projects.

In this group of students there is a wide range of project quality (as measured by elaborateness and creativity). However, this range of quality is about the same between students with artists as parents and those without.

The teacher wants to make sure that they award lollipops in a fair way, no matter whether the students' parents are artists or not.

**Employee awards** A manager at a sales company is deciding which of their 100 employees should receive each of 10 mid-year awards. The manager's goal is to give awards to employees who *will* have high net sales at the end of the year.

The manager uses the following criteria to decide who should get an award:

- Recent performance reviews
- Mid-year net sales
- Number of years on the job

About 50% of the employees are men, and 50% are women.

In the past, men have achieved higher end-of-year net sales than women.

In this group of employees, there is a wide range of qualifications (as measured by performance reviews, mid-year net sales, and number of years on the job). However, this range of qualifications is about the same between male and female employees.

The manager wants to make sure that this awards process is fair to the employees, no matter their gender.

**Hiring** A hiring manager at a new sales company is reviewing 100 new job applications. Each applicant has submitted a resume, and has had an interview. The manager will send job offers to 10 out of the 100 applicants. Their goal is to make offers to applicants who will have high net sales after a year on the job.

The manager will use the following to decide which applicants should receive job offers:

- Interview scores
- Quality of recommendation letters
- Number of years of prior experience in the field

About 50% of the applicants are men, and 50% are women.

In the past, men have achieved higher net sales than women, after one year on the job.

In this applicant pool there is a wide range of applicant quality (as measured by interview scores, recommendation letters, and years of prior experience in the field). However, the range of quality is about the same for both male and female applicants.

The hiring manager wants to make sure that this hiring process is fair to applicants, no matter their gender.

### Questions

1. To what extent do you agree with the following statement: a scenario similar to the one described above might occur in real life.
  - Strongly agree
  - Agree
  - Neither agree nor disagree
  - Disagree
  - Strongly Disagree
2. How much effort should the <teacher, manager, hiring manager> put in to make sure this decision is fair?  
[short answer - number of hours]

#### C.1.2. RULE DESCRIPTIONS AND QUESTIONS

Unless otherwise noted the rule description is shown above each of the questions for reference. Correct answers are noted in red.

**Art project** The teacher uses the following award rule to distribute lollipops: *The fraction of students who receive lollipops that have artist parents should equal the fraction of students in the class that have artist parents. Similarly, the fraction of students who receive lollipops that do not have artist parents should equal the fraction of students in the class that do not have artist parents.*

Example 1: If 10 out of the 20 students in the class have artist parents, then 2 out of the 4 lollipops would be awarded to students with artist parents (and the remaining 2 would be awarded to students without artist parents).

Example 2: If 5 out of the 20 students in the class have artist parents, then 1 out of the 4 lollipops would be awarded to students with artist parents (and the remaining 3 would be awarded to students without artist parents).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality

of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above award rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

3. Suppose a different teacher is considering awarding lollipops to the whole 4th grade. There are 100 students with artist parents, and 200 students without artist parents. The teacher decides to award 10 lollipops to students with artist parents. **Assuming the teacher is required to use the award rule above**, how many students without artist parents need to receive lollipops?

- (a) 10
- (b) 20
- (c) 40
- (d) 50

4. **Assuming the teacher is required to use the award rule above**, in which of these cases can a teacher award more lollipops to students without artist parents than to students with artist parents?

- (a) When the students without artist parents have higher-quality projects (i.e., more elaborate and more creative) than those with artist parents.
- (b) **When there are more students without artist parents than those with artist parents.**
- (c) When students without artist parents have more creative projects than those with artist parents.
- (d) This cannot happen under the award rule.

5. **Assuming the teacher is required to use the award rule above**, is the following statement **TRUE OR FALSE**: Even if a student with artist parents has a project that is of the same quality (i.e., equally elaborate and equally creative) as another project by a student without artist parents, they can be treated differently (i.e., only one of the students might get a lollipop).

6. **Assuming the teacher is required to use the award rule above**, is the following statement **TRUE OR FALSE**: If all students without artist parents have low-quality projects (i.e., low elaborateness and low creativity), but the teacher awards lollipops to some of them, then any lollipops awarded to students with artist parents must be awarded to those who have low-quality projects.

7. **Assuming the teacher is required to use the award rule above**, is the following statement **TRUE OR FALSE**: Suppose the teacher is distributing 10 lollipops amongst a pool of students that includes students with

and without artist parents. Even if all students with artist parents have low-quality (i.e., low elaborateness and low creativity) projects, some of them must still receive lollipops.

8. **Assuming the teacher is required to use the award rule above**, is the following statement **TRUE OR FALSE**: This award rule always allows the teacher to award lollipops exclusively to the students who have the highest quality (i.e., most elaborate and most creative) projects.

In the two examples above there are 20 students. Consider a different scenario, with **6 students – 4 with artist parents and 2 without, as illustrated below**. The next three questions each give a potential outcome for all six students (i.e., which of the 6 students receive awards). Please indicate which of the outcomes follow **the award rule above**.



9. Alternative scenario 1:



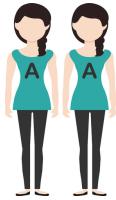
Does this distribution of awards obey the **award rule**?  
**Yes**

10. Alternative scenario 2:



Does this distribution of awards obey the **award rule**?  
**No**

11. Alternative scenario 3:



Does this distribution of awards obey the **award rule**?

No

12. In your own words, explain the **award rule**. [short answer] (The rule is not shown above this question)
13. To what extent do you agree with the following statement: I am confident I know how to **apply the award rule described above**?
  - Strongly agree
  - Agree
  - Neither agree nor disagree
  - Disagree
  - Strongly Disagree
14. Please select the choice that best describes your experience: When I answered the previous questions...
  - (a) I applied the provided award rule only.
  - (b) I used my own ideas of what the correct award decision should be rather than the provided award rule.
  - (c) I used a combination of the provided award rule and my own ideas of what the correct award decision should be.
15. What is your opinion on the award rule? Please explain why. [short answer]
16. Suppose that you are the teacher whose job it is to distribute lollipops to students based on the criteria listed above (i.e., elaborateness of each project, creativity of each project). How would you ensure that this process is fair? [short answer]
17. Was there anything about this survey that was hard to understand or answer? [short answer]

**Employee awards** The manager uses the following award rule to distribute awards: *The fraction of employees who receive awards that are female should equal the fraction of employees that are female. Similarly, fraction of employees who receive awards that are male should equal the fraction of employees that are male.*

Example 1: If there are 50 female employees out of 100, then 5 out of the 10 awards should be awarded to female employees (and the remaining 5 would be made to male employees).

Example 2: If there are 30 female employees out of 100, then 3 out of the 10 awards should be awarded to female employees (and the remaining 7 would be made to male employees).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above award rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

3. Suppose a different manager is considering employees for a different award. There are 100 male employees and 200 female employees, and they decide to give awards to 10 male employees. **Assuming the manager is required to use the award rule above**, how many female employees do they need to give awards to?
  - (a) 10
  - (b) **20**
  - (c) 40
  - (d) 50
4. **Assuming the manager is required to use the award rule above**, in which of these cases can a manager give more awards to female employees than to male employees?
  - (a) When there are more well-qualified female employees than well-qualified male employees (i.e., more women have better performance reviews, higher mid-year net sales, and more years on the job).
  - (b) **When there are more female employees than male employees.**
  - (c) When female employees receive higher performance reviews than male employees.
  - (d) This cannot happen under the award rule.
5. **Assuming the manager is required to use the award rule above**, is the following statement **TRUE OR FALSE**: Even if a male employees qualifications look similar to a female employees (in terms of performance reviews, mid-year net sales, and years on the job), he can be treated differently (i.e., only one of the employees gets an award).
6. **Assuming the manager is required to use the award rule above**, is the following statement **TRUE OR FALSE**: If all female employees are unqualified (i.e.,

have low performance reviews, low mid-year net sales, and few years on the job), but you give awards to some of them, then awards given to male employees must be made to unqualified male employees.

7. **Assuming the manager is required to use the award rule above**, is the following statement **TRUE OR FALSE**: Suppose the manager is distributing 10 awards amongst a pool that includes both male and female employees. Even if all male employees are unqualified for an award (i.e., have low performance reviews, low mid-year net sales, and few years on the job), some of them must still receive awards.
8. **Assuming the manager is required to use the award rule above**, is the following statement **TRUE OR FALSE**: This award rule always allows the manager to distribute awards exclusively to the most qualified employees (i.e., employees with better performance reviews, high mid-year net sales, and high number of years on the job).

In the two examples above there are 100 employees. Consider a different scenario, with **6 employees– 4 female and 2 male, as illustrated below**. The next three questions each give a potential outcome for all six employees (i.e., which of the 6 employees receive awards). Please indicate which of the outcomes follow **the award rule above**.



9. Alternative scenario 1:



Does this distribution of awards obey the **award rule?**  
**Yes**

10. Alternative scenario 2:



Does this distribution of awards obey the **award rule?**  
**No**

11. Alternative scenario 3:



Does this distribution of awards obey the **award rule?**  
**No**

12. In your own words, explain the **award rule**. [short answer] (The rule is not shown above this question)
13. To what extent do you agree with the following statement: I am confident I know how to **apply the award rule described above**?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...
  - (a) I applied the provided award rule only.
  - (b) I used my own ideas of what the correct award decision should be rather than the provided award rule.
  - (c) I used a combination of the provided award rule and my own ideas of what the correct award decision should be.
15. What is your opinion on the award rule? Please explain why. [short answer]
16. Suppose that you are the manager whose job it is to distribute mid-year awards to employees based on the criteria listed above (i.e., recent performance reviews, mid-year net sales, number of years on the job). How would you ensure that this process is fair? [short answer]
17. Was there anything about this survey that was hard to understand or answer? [short answer]

**Hiring** The hiring manager uses the following hiring rule to send out offers: *The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, fraction of applicants who*

*receive job offers that are male should equal the fraction of applicants that are male.*

Example 1: If there are 50 female applicants out of the 100 applicants, then 5 out of the 10 offers would be made to female applicants (and the remaining 5 would be made to male applicants).

Example 2: If there are 30 female applicants out of the 100 applicants, then 3 out of the 10 offers would be made to female applicants (and the remaining 7 would be made to male applicants).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

3. Suppose a different hiring manager is considering applicants for a different job. There are 100 male applicants and 200 female applicants, and they decide to send offers to 10 male applicants. **Assuming the hiring manager is required to use the hiring rule above**, how many female applicants do they need to send offers to?

- (a) 10
- (b) **20**
- (c) 40
- (d) 50

4. **Assuming the hiring manager is required to use the hiring rule above**, in which of these cases can a hiring manager make more job offers to female applicants than to male applicants?

- (a) When there are more well-qualified female applicants than well-qualified male applicants (i.e., more women have higher interview scores, higher quality recommendation letters, and more years of prior experience in the field).
- (b) **When there are more female applicants than male applicants.**
- (c) When female applicants receive better interview scores than male applicants.
- (d) This cannot happen under the hiring rule.

5. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement **TRUE** OR FALSE: Even if a male applicants qualifications

look similar to a female applicants (in terms of interview scores, recommendation letters, and years of prior experience in the field), he can be treated differently (i.e., only one of the applicants will receive a job offer).

6. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement **TRUE** OR **FALSE**: If all female applicants are unqualified (i.e., have low interview scores, low-quality recommendation letters, and few years of prior experience in the field), but you send job offers to some of them, then any job offers made to male applicants must be made to unqualified male applicants.

7. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement **TRUE** OR **FALSE**: Suppose the hiring manager is sending out 10 job offers to a pool that includes male and female applicants. Even if all male applicants are unqualified (i.e., have low interview scores, low-quality recommendation letters, and few years of prior experience in the field), some of them must still receive job offers.

8. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement **TRUE** OR **FALSE**: This hiring rule always allows the hiring manager to send offers exclusively to the most qualified applicants (i.e., applicants with high interview scores, high quality recommendation letters, and high number years of prior experience in the field).

In the two examples above there are 100 applicants. Consider a different scenario, with **6 applicants – 4 female and 2 male**, as illustrated below. The next three questions each give a potential outcome for all 6 applicants (i.e., which of the 6 applicants receive job offers). Please indicate which of the outcomes follow **the hiring rule above**.



9. Alternative scenario 1:



Does this distribution of job offers obey the **hiring rule**? **Yes**

10. Alternative scenario 2:



Does this distribution of job offers obey the **hiring rule**? **No**

11. Alternative scenario 3:



Does this distribution of job offers obey the **hiring rule**? **No**

12. In your own words, explain the **hiring rule**. [short answer] (The rule is not shown above this question)

13. To what extent do you agree with the following statement: I am confident I know how to **apply the hiring rule described above?**

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

- I applied the provided hiring rule only.
- I used my own ideas of what the correct hiring decision should be rather than the provided hiring rule.
- I used a combination of the provided hiring rule and my own ideas of what the correct hiring decision should be.

15. What is your opinion on the hiring rule? Please explain why. [short answer]

16. Suppose that you are the hiring manager whose job it is to send job offers to applicants based on the criteria

listed above (i.e., interview scores, quality of recommendation letters, number of years of prior experience in the field). How would you ensure that this process is fair? [short answer]

17. Was there anything about this survey that was hard to understand or answer? [short answer]

## C.2. Study-2: Survey

Each of the surveys are split into four main sections. The first section is the consent form which can be found in Appendix D. The second section describes the hiring scenario and asks questions about it (§C.2.1). The third section describes the fairness metric, defined as the rule, used (in this case it is demographic parity) and asks specific questions about the metric (§C.2.2). Finally the last section asks for demographic information (§C.3).

### C.2.1. SCENARIO DESCRIPTION AND QUESTIONS

The following is shown to each participant (note that Step 3 is not shown to participants with the DP definition):

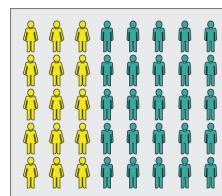
It is very important that you read each question carefully and think about your answers. The success of our research relies on our respondents being thoughtful and taking this task seriously.

I have read the above instructions carefully.

A company, Sales-a-lot, is reviewing their hiring process. They want to hire applicants who are high performing, and they also want to make sure that their hiring process is fair to their applicants, no matter their gender. To do this, Sales-a-lot employs an external firm, Recruit-a-matic, which keeps track of all applicants. This review will take place over one year.

For clarity at each stage of the hiring process we use images to represent the hiring pool.

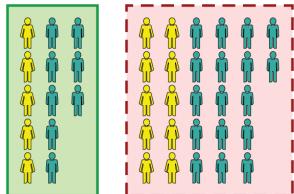
**Step 1: Applicant Pool.** At the beginning of the year, Sales-a-lot reviews all job applicants, and sends job offers to some of them. The initial applicant pool is shown with a gray background. For example, the following image shows an applicant pool with 15 female applicants and 25 male applicants:



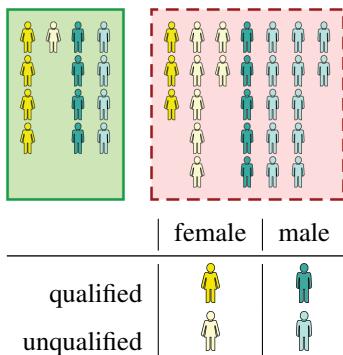
**Step 2: Sending Job Offers.** Next, Sales-a-lot sends job offers to some of these applicants, using the following criteria:

- Interview scores
- Quality of recommendation letters
- Number of years of prior experience in the field

Suppose that Sales-a-lot sends offers to 5 female applicants and 8 male applicants (so 10 female and 17 male applicants didn't receive offers). In the following image, applicants who received a job offer are shown on the left (with a green background) and applicants who didn't receive a job offer are shown on the right, with a red background):



**Step 3: Applicant Evaluation.** For the rest of the year, Recruit-a-matic (the external firm) keeps track of all applicants in the initial pool, whether they received job offers or not. At the end of the year, Recruit-a-matic finds out which applicants were high performers, i.e. qualified (shown in dark), and which applicants were low performers, i.e. unqualified (shown in light). For example, the following image shows that most of the high performers received job offers, but some did not.



### Questions

1. To what extent do you agree with the following statement: a scenario similar to the one described above might occur in real life.
  - Strongly agree
  - Agree
  - Neither agree nor disagree
  - Disagree
  - Strongly disagree
2. How much effort, in hours, should Sales-a-lot put in to make sure these decisions were fair? [short answer - number of hours]

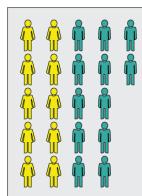
### C.2.2. RULE DESCRIPTIONS AND QUESTIONS

The following sections provide fairness definitions (presented to participants as *rules*) for Demographic Parity, Equal Opportunity (FNR and FPR), and Equalized Odds. Unless otherwise noted the rule description is shown above each of the questions for reference. Correct answers are noted in red.

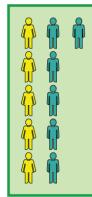
**Demographic Parity.** Recruit-a-matic uses the following rule to determine whether Sales-a-lots hiring decisions were fair:

*The fraction of male candidates who receive job offers should equal the fraction of female candidates who receive job offers.*

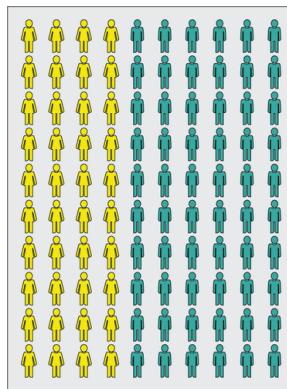
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following applicants (10 female and 12 male).



If Sales-a-lot sent job offers to the following number of applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

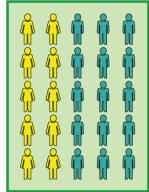


Example 2: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot reviewed a total of 100 applicants as follows (40 female and 60 male).



If Sales-a-lot sent job offers to the following number of

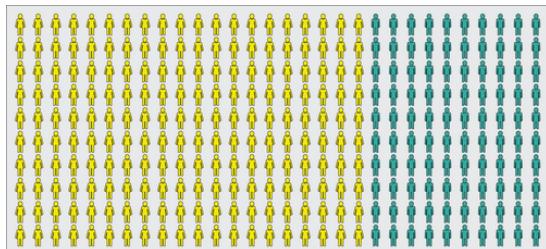
applicants (10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



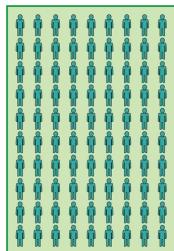
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

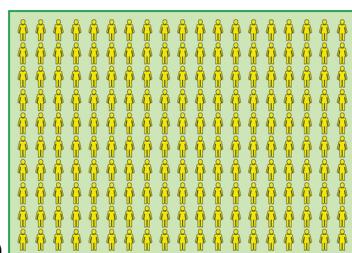
3. Suppose a different company considered applicants for a different job. There were 200 female applicants and 100 male applicants,



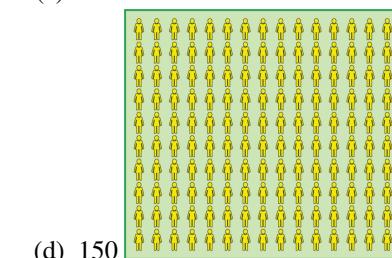
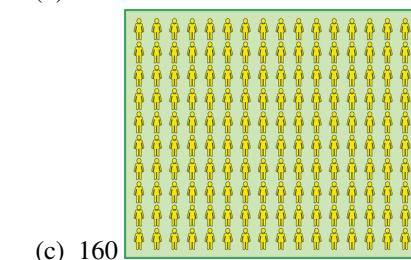
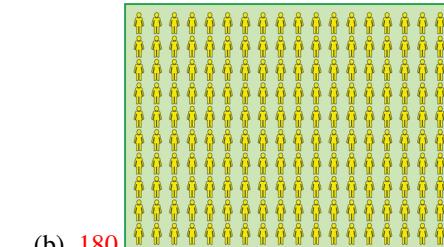
and they did send job offers to 90 male applicants.



Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many female applicants should have received job offers?



(a) 190



4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have accepted more qualified female applicants than qualified male applicants?
- When there are more qualified female applicants than qualified male applicants (i.e., more women had low net sales at the end of the year).
  - When there are more female applicants than male applicants.**
  - When female applicants receive worse interview scores than male applicants.
  - This cannot happen under the hiring rule.
5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement **TRUE OR FALSE**: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.
6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement **TRUE OR FALSE**: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.
7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the

following statement **TRUE** OR **FALSE**: The hiring rule above requires that even if all male applicants were unqualified (they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.

8. Is the following statement **TRUE** OR **FALSE**: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 applicants – 4 female and 2 male, as illustrated below. The next three questions each give a different potential outcome for all 6 applicants (i.e., which of the 6 applicants do receive job offers). Please indicate which of the outcomes follow the hiring rule above.



9. Sales-a-lot makes the following hiring decisions.



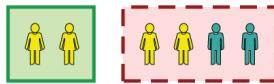
Do these decisions obey the hiring rule? **Yes**

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

- (a) I applied the provided hiring rule only.
- (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.
- (c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

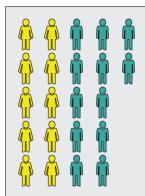
17. Please explain your opinion on the hiring rule. [short answer]

18. Was there anything about this survey that was hard to understand or answer? [short answer]

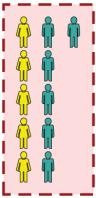
**Equal Opportunity - FNR.** Recruit-a-matic uses the following rule to determine whether Sales-a-lots hiring decisions were fair:

*The fraction of qualified male candidates who do not receive job offers should equal the fraction of qualified female candidates who do not receive job offers.*

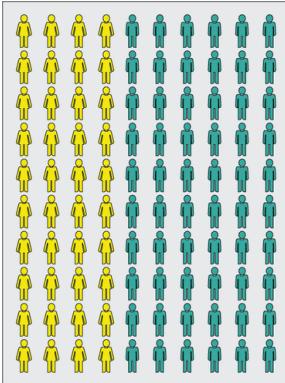
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following qualified applicants (10 female and 12 male).



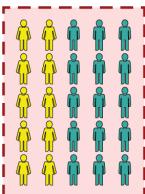
If Sales-a-lot did not send job offers to the following number of qualified applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



Example 2: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot reviewed a total of 100 qualified applicants as follows (40 female and 60 male).



If Sales-a-lot did not send job offers to the following number of qualified applicants (10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

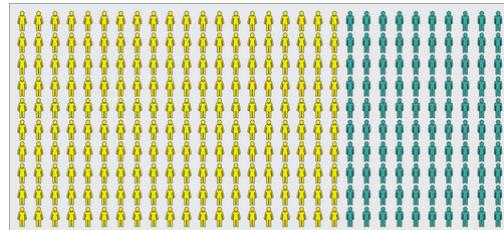


Note that in the above examples the remaining qualified applicants received job offers, but are not displayed here.

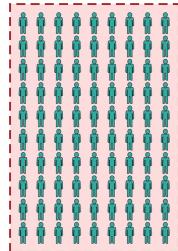
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

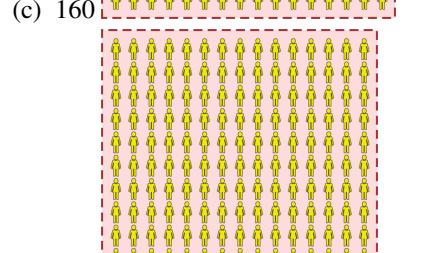
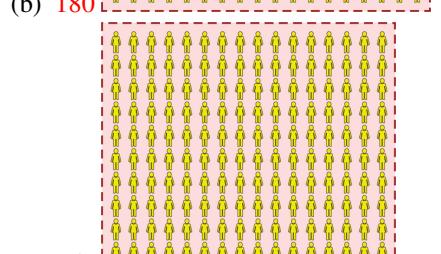
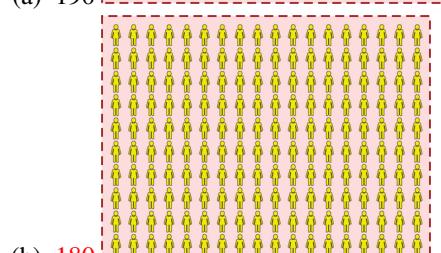
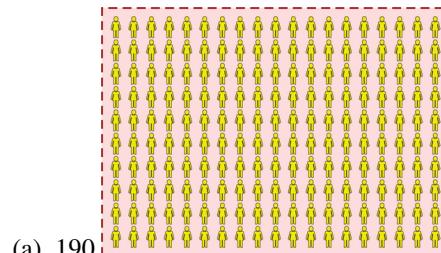
3. Suppose a different company considered applicants for a different job. There were 200 qualified female applicants and 100 qualified male applicants,



and they did not send job offers to 90 qualified male applicants.

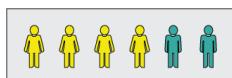


Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many qualified female applicants should not have received job offers?



4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have rejected more qualified female applicants than qualified male applicants?
  - (a) When there are more qualified female applicants than qualified male applicants (i.e., more women had low net sales at the end of the year).
  - (b) When there are more female applicants than male applicants.
  - (c) When female applicants receive worse interview scores than male applicants.
  - (d) This cannot happen under the hiring rule.
  
5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement **TRUE OR FALSE**: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.
  
6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement **TRUE OR FALSE**: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.
  
7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement **TRUE OR FALSE**: The hiring rule above requires that even if all male applicants were unqualified (they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.
  
8. Is the following statement **TRUE OR FALSE**: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 qualified applicants – 4 female and 2 male, as illustrated below. The next three questions each give a different potential outcome for all 6 qualified applicants (i.e., which of the 6 applicants do not receive job offers). Please indicate which of the outcomes follow the hiring rule above.



9. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **Yes**

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

- (a) I applied the provided hiring rule only.
- (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.
- (c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?

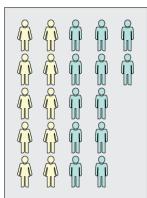
- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

17. Please explain your opinion on the hiring rule. [short answer]
18. Was there anything about this survey that was hard to understand or answer? [short answer]

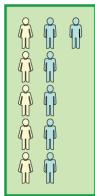
**Equal Opportunity - FPR.** Recruit-a-matic uses the following rule to determine whether Sales-a-lots hiring decisions were fair:

*The fraction of unqualified male candidates who receive job offers should equal the fraction of unqualified female candidates who receive job offers.*

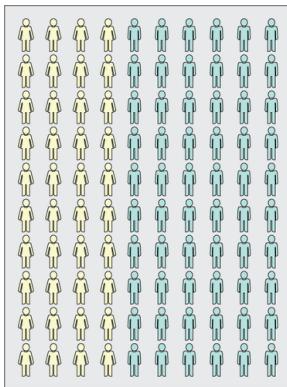
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following unqualified applicants (10 female and 12 male).



If Sales-a-lot sent job offers to the following number of unqualified applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

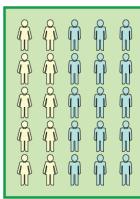


Example 2: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot reviewed a total of 100 unqualified applicants as follows (40 female and 60 male).



If Sales-a-lot sent job offers to the following number of unqualified applicants (10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the

hiring rule).

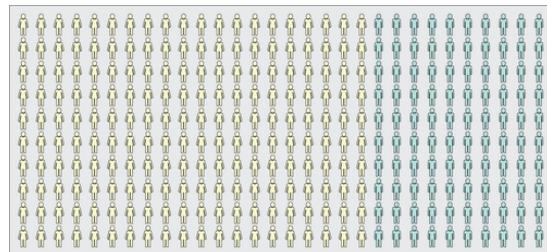


Note that in the above examples the remaining unqualified applicants did not receive job offers, but are not displayed here.

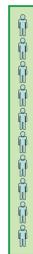
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

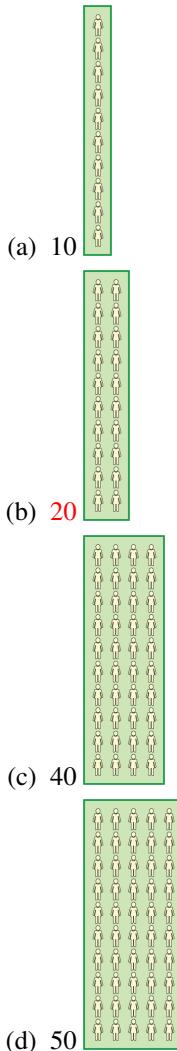
3. Suppose a different company considered applicants for a different job. There were 200 unqualified female applicants and 100 unqualified male applicants,



and they did send job offers to 10 unqualified male applicants.



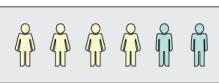
Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many unqualified female applicants should have received job offers?



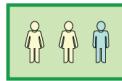
4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have accepted more unqualified female applicants than unqualified male applicants?
- When there are more unqualified female applicants than unqualified male applicants (i.e., more women had low net sales at the end of the year).**
  - When there are more female applicants than male applicants.
  - When female applicants receive worse interview scores than male applicants.
  - This cannot happen under the hiring rule.
5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement **TRUE OR FALSE**: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.

- Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement **TRUE OR FALSE**: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.
- Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement **TRUE OR FALSE**: The hiring rule above requires that even if all male applicants were unqualified (they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.
- Is the following statement **TRUE OR FALSE**: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 unqualified applicants – 4 female and 2 male, as illustrated below. The next three questions each give a different potential outcome for all 6 applicants (i.e., which of the 6 applicants receive job offers). Please indicate which of the outcomes follow the hiring rule above.

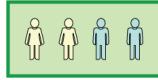


9. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **Yes**

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

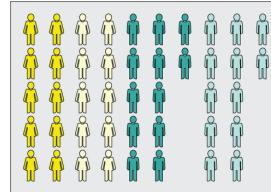
11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

- In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]
- To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree



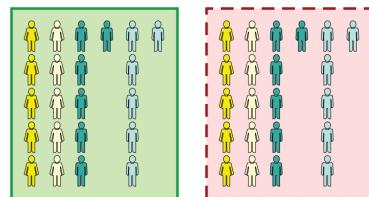
14. Please select the choice that best describes your experience: When I answered the previous questions...

- I applied the provided hiring rule only.
- I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.
- I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

If Sales-a-lot did send offers to the following number of unqualified applicants (left, 5 female and 6 male), and did not send job offers to the following number of qualified applicants (right, 5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



16. To what extent do you agree with the following statement: I agree with the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

17. Please explain your opinion on the hiring rule. [short answer]

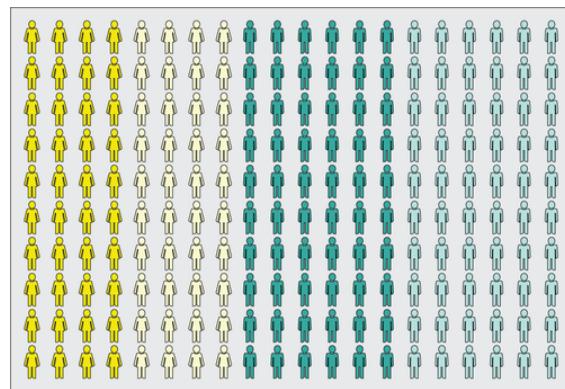
18. Was there anything about this survey that was hard to understand or answer? [short answer]

**Equalized Odds.** Recruit-a-matic uses the following rule to determine whether Sales-a-lots hiring decisions were fair:

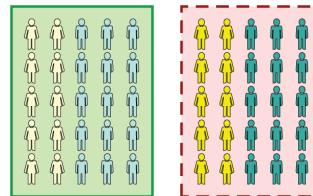
*The fraction of qualified male candidates who do not receive job offers should equal the fraction of qualified female candidates who do not receive job offers. Similarly, the fraction of unqualified male candidates who receive job offers should equal the fraction of unqualified female candidates who receive job offers.*

Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following qualified applicants (10 female and 12 male) and unqualified applicants (10 female and 12 male).

Example 2: Suppose that over the past year, Recruit-a-lot finds that Sales-a-lot reviewed a total of 100 qualified applicants (40 female and 60 male) and 100 unqualified applicants (40 female and 60 male).



If Sales-a-lot did send offers to the following number of unqualified applicants (left, 10 female and 15 male), and did not send job offers to the following number of qualified applicants (right, 10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



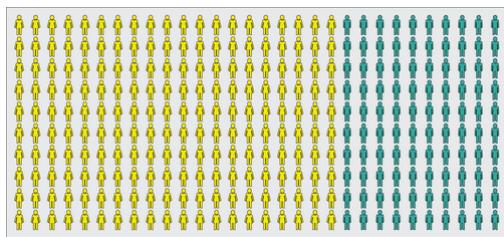
Note that in the above examples the remaining unqualified applicants did not receive job offers, but are not displayed here. Similarly, the remaining qualified applicants received

job offers, but are not displayed here.

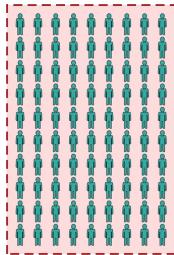
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

**Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.**

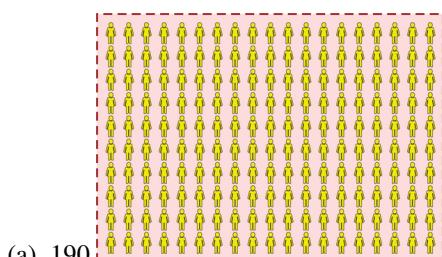
3. Suppose a different company considered applicants for a different job. There were 200 qualified female applicants and 100 qualified male applicants,



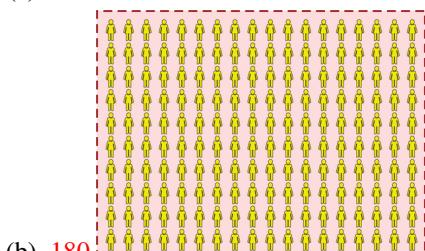
and they did not send job offers to 90 qualified male applicants.



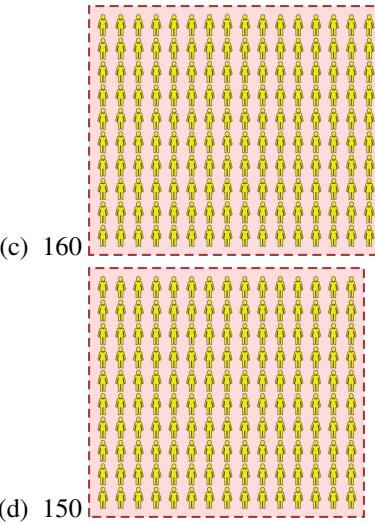
Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many qualified female applicants should not have received job offers?



(a) 190



(b) 180



4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have accepted more unqualified female applicants than unqualified male applicants?

- (a) When there are more unqualified female applicants than unqualified male applicants (i.e., more women had low net sales at the end of the year).
- (b) When there are more female applicants than male applicants.
- (c) When female applicants receive worse interview scores than male applicants.
- (d) This cannot happen under the hiring rule.

5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement **TRUE** OR **FALSE**: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.

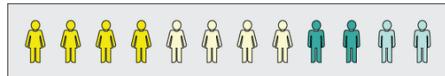
6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement **TRUE** OR **FALSE**: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.

7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement **TRUE** OR **FALSE**: The hiring rule above requires that even if all male applicants were unqualified (they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.

8. Is the following statement **TRUE** OR **FALSE**: The hiring rule above always allows Sales-a-lot to send job

offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 qualified applicants – 4 female and 2 male; and 6 unqualified applicants – 4 female and 2 male. The next three questions each give a different potential outcome for the applicants (i.e., which of the applicants did or did not receive job offers). Please indicate which of the outcomes follow the hiring rule above.



9. Sales-a-lot makes the following hiring decisions.



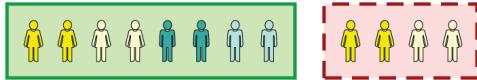
Do these decisions obey the hiring rule? **Yes**

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? **No**

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

- (a) I applied the provided hiring rule only.
- (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.
- (c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly Disagree

17. Please explain your opinion on the hiring rule. [short answer]

18. Was there anything about this survey that was hard to understand or answer? [short answer]

### C.3. Demographic Information

1. Please specify the gender with which you most closely identify:

- Male
- Female
- Other
- Prefer not to answer

2. Please specify your year of birth

3. Please specify your ethnicity (you may select more than one):

- White
- Hispanic or Latinx
- Black or African American
- American Indian or Alaska Native
- Asian, Native Hawaiian, or Pacific Islander
- Other

4. Please specify the highest degree or level of school you have completed:

- Some high school credit, no diploma or equivalent
- High school graduate, diploma or the equivalent (for example: GED)
- Some college credit, no degree
- Trade/technical/vocational training
- Associates degree
- Bachelors degree

- Masters degree
  - Professional or doctoral degree (JD, MD, PhD)
5. How much experience do you have in each of the following areas? (1 - no experience, 2 - limited experience, 3 - significant experience, 4 - expert)
- (a) Human resources (making hiring decisions)
  - (b) Management (of employees)
  - (c) Education (teaching)
  - (d) IT infrastructure/systems administration
  - (e) Computer science/programming
  - (f) Machine learning/data science

**We will maintain privacy of the information you have provided here. Your information will only be used for data analysis purposes.**

## **D. Consent**

### **D.1. Online Survey Consent Form**

#### **D.1.1. PROJECT TITLE**

Fairness Evaluation and Comprehension

#### **D.1.2. PURPOSE OF THE STUDY**

This research is being conducted by [Blinded] at [Blinded]. We are inviting you to participate in this research project because you are above 18. The purpose of this research project is to understand lay comprehension of different fairness metrics.

#### **D.1.3. PROCEDURES**

The procedures will start with reading a brief description of a decision-making scenario. You will then be asked to answer some comprehension questions about the scenario. The questions will look like the following: What are the pros and cons of the notion of fairness described above?

Finally, you will be asked some demographics questions. The entire survey will take approximately 20 minutes or less.

#### **D.1.4. POTENTIAL RISKS AND DISCOMFORTS**

There are several questions to answer over the course of this study, so you may find yourself growing tired towards the end. Outside of this, there are minimal risks to participating in this research study. All data collected in this study will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.

However, if at any time you feel that you wish to terminate your participation for any reason, you are permitted to do so.

#### **D.1.5. POTENTIAL BENEFITS**

There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of fairness metrics and their applications.

#### **D.1.6. CONFIDENTIALITY**

Any potential loss of confidentiality will be minimized by storing all data (including information such as MTurk IDs and demographics) will be stored securely (a) in a password-protected computer located at [Blinded] or (b) using a trusted third party (Qualtrics). Personally identifiable information that is collected (MTurk IDs, IP addresses, cookies) will be deleted upon study completion. All other data gathered will be stored for three years post study completion, after which it will be erased. The only persons that will have access to the data are the Principle Investigator and the Co-Investigators.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the [Blinded] or governmental authorities if you or someone else is in danger or if we are required to do so by law.

#### **D.1.7. COMPENSATION**

You will receive \$3. You will be responsible for any taxes assessed on the compensation.

If you will earn \$100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over \$100 only your name and address will be collected to receive compensation.

#### **D.1.8. RIGHT TO WITHDRAW AND QUESTIONS**

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator: [Blinded]

#### **D.1.9. PARTICIPANT RIGHTS**

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

[Blinded]

For more information regarding participant rights, please visit: [Blinded]

This research has been reviewed according to the [Blinded] IRB procedures for research involving human subjects.

#### D.1.10. STATEMENT OF CONSENT

By agreeing below you indicate that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. Please ensure you have made a copy of the above consent form for your records.

Please ensure you have made a copy of the above consent form for your records. A copy of this consent form can be found here [link to digital copy].

- I am age 18 or older
- I have read this consent form
- I voluntarily agree to participate in this research study

### D.2. Cognitive Interview Consent Form

#### D.2.1. PROJECT TITLE

Fairness Cognitive Interview

#### D.2.2. PURPOSE OF THE STUDY

This research is being conducted by [Blinded] at [Blinded]. We are inviting you to participate in this research project because you are above the age of 18, and fluent in English. The purpose of this research project is to understand lay comprehension of different fairness metrics.

#### D.2.3. PROCEDURES

The procedure involves completing an interview. The full procedure will be approximately 1 hour in duration.

During the interview you will be audio recorded, if you agree to be recorded. You will be asked to first read a brief description of a decision-making scenario. You will then be asked to fill out a survey about the scenario. While answering questions you will be asked verbal questions related to how you reached your answer in the survey.

Sample survey question: Is the following statement true or false? This hiring rule allows the hiring manager to send offers exclusively to the most qualified applicants.

Sample interview question: How did you reach your answer to that survey question?

#### D.2.4. POTENTIAL RISKS AND DISCOMFORTS

There are several questions to answer over the course of this study, so you may find yourself growing tired towards the end. Outside of this, there are minimal risks to participating in this research study. All data collected in this study will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.

However, if at any time you feel that you wish to terminate your participation for any reason, you are permitted to do so.

#### D.2.5. POTENTIAL BENEFITS

There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of fairness metrics and their applications.

#### D.2.6. CONFIDENTIALITY

Any potential loss of confidentiality will be minimized by storing all data (including information such as demographics) securely (a) in a password protected computer located at [Blinded] or (b) using a trusted third party (Qualtrics). Personally identifiable information that is collected will be deleted upon study completion. All other data gathered will be stored for three years post study completion, after which it will be erased. The only persons that will have access to the data are the principle Investigator and the Co-Investigators.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of [Blinded] or governmental authorities if you or someone else is in danger or if we are required to do so by law.

#### D.2.7. COMPENSATION

You will receive \$30. You will be responsible for any taxes assessed on the compensation.

If you will earn \$100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over \$100 only your name and address will be collected to receive compensation.

#### D.2.8. RIGHT TO WITHDRAW AND QUESTIONS

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized

or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator: [Blinded]

#### D.2.9. PARTICIPANT RIGHTS

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact: [Blinded]

For more information regarding participant rights, please visit: [Blinded]

This research has been reviewed according to [Blinded] IRB procedures for research involving human subjects.

#### D.2.10. STATEMENT OF CONSENT

Your signature indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. You will receive a copy of this signed consent form.

Please initial all that apply (you may choose any number of these statements):

- I agree to be audio recorded
- I agree to allow researchers to use my audio recording in research publications and presentations.
- I do not agree to be audio recorded

If you agree to participate, please sign your name below.

## E. Methods

### E.1. Cognitive Interviews

We recruited 9 participants from a large metropolitan area using Craigslist. We required participants to be over 18 years of age and fluent in English. Participants ranged between the ages of 20 and 66. These interviews took place on our organization’s campus and lasted about 1 hour. All participants signed a written consent form prior to the interview, and were paid \$30 for their time.

During these interviews, participants completed a preliminary version of the survey used in Study-1. After each survey question, we asked the participants several interview questions related to their comprehension of and feelings toward the survey. We found that some participants tended to use their own personal notions of fairness when answering comprehension questions rather than using the definition

we provided. We were concerned that this would limit our ability to effectively measure comprehension. To address this problem, we rewrote several parts of our survey and added two new questions (Q14 and Q15).

### E.2. Non-Expert Verification

We designed this study to assess *non-expert* understanding and opinions of ML fairness metrics. To this end, we asked respondents to self-rate their level of expertise in a variety of fields, including ML, at the end of the survey (see Appendix C.3). A number of participants did report having “expert” level experience in ML ( $n = 2$  out of 147 in Study-1, and  $n = 15$  out of 349 in Study-2). We considered removing these participants from the analyses, but ultimately did not because there was no relationship between self-reported ML expertise and comprehension score (Spearman’s rho, for both studies).