Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics

Mallak Alkhathlan
Worcester Polytechnic Institute
Worcester, Massachussetts, USA
malkhatlan@wpi.edu

Kathleen Cachel
Worcester Polytechnic Institute
Worcester, Massachussetts, USA
kcachel@wpi.edu

Hilson Shrestha
Worcester Polytechnic Institute
Worcester, Massachussetts, USA
hshrestha@wpi.edu

Lane Harrison
Worcester Polytechnic Institute
Worcester, Massachussetts, USA
ltharrison@wpi.edu

Elke Rundensteiner
Worcester Polytechnic Institute
Worcester, Massachussetts, USA
rundenst@wpi.edu

ABSTRACT

Algorithmic decision-making using rankings— prevalent in areas from hiring and bail to university admissions— raises concerns of potential bias. In this paper, we explore the alignment between people's perceptions of fairness and two popular fairness metrics designed for rankings. In a crowdsourced experiment with 480 participants, people rated the perceived fairness of a hypothetical scholarship distribution scenario. Results suggest a strong inclination towards relying on explicit score values. There is also evidence of people's preference for one fairness metric, NDKL, over the other metric, ARP. Qualitative results paint a more complex picture: some participants endorse meritocratic award schemes and express concerns about fairness metrics being used to modify rankings; while other participants acknowledge socio-economic factors in score-based rankings as justification for adjusting rankings. In summary, we find that operationalizing algorithmic fairness in practice is a balancing act between mitigating harms towards marginalized groups and societal conventions of leveraging traditional performance scores such as grades in decision-making contexts.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

Algorithmic Fairness, Human Perception of Fairness

ACM Reference Format:
Mallak Alkhathlan, Kathleen Cachel, Hilson Shrestha, Lane Harrison, and Elke Rundensteiner. 2024. Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3630106.3659018

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659018

1 INTRODUCTION

As artificial intelligence (AI) and automated decision-making systems critically impact more of our daily lives, there is an increasing need to ensure that these technologies do not disproportionately harm or replicate societal bias toward disadvantaged populations and legally protected groups [8, 38, 40, 42, 56, 69, 72]. The algorithmic fairness community has developed metrics that conceptualize, measure, and mathematically formulate various definitions of fairness. Practitioners ranging from data scientists and researchers to regulators and auditors use such metrics to assess the fairness harms of technologies. Moreover, these metrics are operationalized in algorithms to create fairness-aware methods.

When developing AI systems, fairness metrics can play a central role in the mitigation of harm toward marginalized groups. While fairness metrics are mathematical formulas, they are also inherently sociotechnical constructs, translating societal notions into numeric scores. Several recent efforts highlight the challenge of transcribing human values into precise equations [4, 24, 25, 29, 60, 61, 70]. Fairness metrics have been formulated for specific problem settings such as classification, selection, rankings, etc. While much research has focused on fair classification [4, 24, 25, 29, 60, 61, 70] and recommender systems [58, 66], significant gaps in understanding fairness in rankings still persist. These alternate problem settings pose open questions that remain largely unexplored. Not only do fair ranking metrics differ from fair classification metrics, rankings are also inherently complex objects with object placement dependencies, namely, moving someone up in a ranking also implies someone else is moved down.

In this paper, our aim is to explore the alignment between contemporary fair rankings metrics and what people believe is a fair outcome. We design an experiment context around student scholarship allocation— modeled as a group fair ranking problem. A large body of fair-ranking research focuses on ensuring groups receive comparable shares of favorable outcomes in a ranking [65, 79, 81], which is referred to as the fairness construct of statistical parity [18]. Statistical parity has been formulated into a number of fair ranking metrics. We employ two of the recently proposed metrics, Normalized Discounted Cumulative KL-divergence (NDKL) [21] and Attribute Rank Parity [13]. We construct a series of scenarios presenting students and their performance data (grades), along with a scholarship ranking ordering students for receiving decreasing amounts of award funding. The scholarship rankings

arrange students by how these two sometimes conflicting metrics conceptualize and quantify fairness.

We conduct a controlled crowd-sourced experiment asking participants to rate the fairness of each student scholarship ranking. Our investigation is steered by three research questions:

- (RQ1) How do different fair ranking metrics align with people's perceptions of fairness?
- (RQ2) Does the inclusion of demographic information, such as race, influence people's perception of these fairness metrics?
- (RQ3) Do people's perceptions of fairness change when candidates being ranked are similar in their score-based performance (such as grades of students) versus dissimilar?

We assess the fairness of scholarship distribution within a fictional public school district, examining how committee-determined factors like grades and socio-economic backgrounds influence perceptions of fairness across several conditions (see Section 3). First, findings from the *quantitative study* strongly suggest a preference for orderings that closely align with an ordering based on performance scores (grade) ***(RQ1)***. We observed this preference across all factors. Second, introducing demographic distinctions, such as black and white groups, influenced the perception of fairness in unbalanced groups (2 vs. 6 and 6 vs. 2 splits) ***(RQ2)***. Third, we found that people's perceptions of fairness are more consistent when evaluating candidates with 'similar grades' as opposed to scenarios with 'dissimilar grades'. This observation remains true in situations when the groups are of equal size '(4 vs. 4)'. We noticed this pattern in both 'concrete' considerations, such as 'black vs. white', and 'abstract' groups, like 'yellow vs. purple' ***(RQ3)***.

Two main themes emerge from the *qualitative analysis* of free-form participant comments. First, there was a sense of "merit-based discontent" among participants due to perceived inconsistencies in the correlation between grades and ranking orders. Second, while participants acknowledged the importance of "socio-economic factors" in decision-making, they expressed a desire for more transparency when these group-related factors were influencing the final ranking. This concern led some participants to question the efficacy of current fairness-driven ranking methodologies. Additional materials linked to this paper, including survey questions, overview of the experimental design, data and metadata artificats, and code for statistical analysis, are available as supplementary materials[1].

2 RELATED WORK

Numerous metrics exist in the literature aiming for fairness in algorithms, yet it remains unclear which metric best aligns with people's perceptions. Our research explores this area, focusing on the intersection of algorithmic fairness and public perception, particularly with respect to the impacts of racial categorization and education.

2.1 Algorithmic Fairness in Ranking

Algorithmic fairness is broadly divided into two main categories: individual fairness which seeks for similar individuals to receive similar outcomes [18], and group fairness which aims to ensure that protected groups of people are treated comparably [55]. Arguably, the most widely adopted notion of group fairness in ranking-based tasks is statistical parity [36]. Several works propose algorithmic techniques for ensuring statistical parity, typically in the setting of two protected groups [37, 51, 65, 79, 81]. We focus on metrics that assess group fairness in rankings, specifically, the Normalized Discounted Cumulative KL-divergence (NDKL) [21] and Attribute Rank Parity (ARP) [13] metrics. While both capture the fairness concept of statistical parity, we chose these two metrics for this study based on their difference in priority given to groups at different parts of the ranking. Further details on these metrics are in Appendix A.

NDKL assesses the representation of different groups at every prefix of the ranking, weighting the higher-up prefixes more. It deems a ranking fair if each prefix, i.e, a top-k set, of the ranking has a proportional share of all groups. NDKL is most fair at 0. It conceptually focused on representing groups fairly higher up in the ranking. In contrast, ARP is a pairwise metric, that decomposes the ranking into pairwise comparisons with a mixed pair comparing candidates of disjoint groups. It measures the difference between average mixed pairs won by each group. It is most fair at 0. It is conceptually geared toward ensuring fair group treatment equally across the entire ranking.

2.2 People's Perceptions and Algorithmic Fairness

Empirical studies [14, 60, 61, 70, 83] have found limited evidence on diverse perceptions of theoretical fairness concepts in algorithms, often focusing on context and stakeholders [41, 71, 77]. Thus they highlight a disconnect between theory and practice. Although there has been significant progress in developing models for fairness distribution [20, 45, 73], there remains a lack of consensus on the most effective fairness notion. Also, these notions, while explored [18, 39, 73], face challenges in terms of their real-world effectiveness and acceptability.

HCI research has focused on fairness in classification [60, 61, 70], yet the public's perception of ranking metrics in algorithmic systems remains less explored. Srivastava's work [70] investigates how mathematical fairness interpretations in binary classification align with human perceptions across various societal contexts. Saxena et al.'s study [61] on fairness perceptions in loan decisions, especially considering applicant race, reveals an integration of quantitative metrics with historical and social contexts. This intersection raises critical questions about the alignment of standard fairness definitions with actual human perceptions. Debjani et al. [60] developed a metric assessing non-experts' understanding of ML fairness, linking it to sentiment and demographics, thereby highlighting the necessity of aligning technical ML fairness models with public perceptions of fairness, with a focus on ethics and trust in machine decision-making.

The exploration of fairness in algorithmic recommendation systems involves two primary stakeholders: providers and consumers. Studies such as [33, 68] focus on consumers' fairness perceptions, while [28, 64] examine specific user groups on TikTok. On the other hand, [63] investigates providers' challenges with online algorithms, focusing on how these algorithms impact the visibility

[1] https://osf.io/bdnq2/

and success of their work. They explore the specific strategies that content creators, particularly romance novelists, employ to adapt to and navigate the changing algorithmic landscapes of digital publishing platforms. Both [17, 67] address fairness between providers and consumers in music streaming and microlending, emphasizing the need to understand all stakeholder perspectives. Our study aims to build on these concepts, emphasizing provider-side fairness in ranking to align peoples' perceptions with fairness metrics for a more inclusive understanding of fairness in digital spaces.

2.3 Racial Inequality in HCI Studies

Recent FAccT community discussions highlight race and fairness as "contested" constructs, questioning their quantifiability [1, 7, 10, 27, 34, 82]. This has shifted focus to real-world algorithmic fairness over theoretical abstraction [6, 19, 26, 84]. Systemic discrimination against Black individuals has led to disparities in resource access, economic exploitation, and persistent stereotypes [2, 30, 33, 76, 78].

Algorithmic tools in policing, such as bail sentencing and police stops, have been found to reinforce biases against black people [3, 43, 44, 52, 57]. Similarly, search engines perpetuate biases and stereotypes, often overlooking the specific issue of anti-Blackness [9, 23, 35, 53]. In healthcare, racial biases in medical algorithms result in the underrepresentation of black patients in high-risk care [22, 54, 74]. Black women face unique challenges such as "algorithmic misogynoir" on social media [28, 31, 50]. These biases are often exacerbated by major companies like Google, Amazon, and Facebook, which amplify racial disparities, especially affecting black communities [5, 59, 62]. Our research examines the impact of racial factors on algorithmic fairness perceptions, particularly how limited access to tutoring, mental health services, and study time present challenges and may affect public perception of fair treatment for some groups [1].

3 STUDY DESIGN AND METHODOLOGY

Key Variables and Conditions

To investigate people's perceived notions of fairness in ranking scenarios, we design an experimental framework that enables the manipulation of variables of interest, including fairness metrics and candidate characteristics:

- **Fairness Metrics:** We examine two ranking-focused fairness metrics, namely, Normalized Discounted Cumulative KL-divergence (NDKL) [21] and Attribute Rank Parity (ARP) [13]. We generate a series of rankings with different combinations of "fair" and "unfair" according to these two metric scores, respectively. For example, fair according to NDKL and unfair according to ARP, etc., see Table 1.
- **Candidate Sensitivity Considerations:** Fairness perceptions are examined under two sets of conditions: *Concrete* and *Abstract*. In *Concrete*, we examine scenarios where candidates are categorized in racially labeled groups—"advantage black" and "advantage white". Meanwhile, in *Abstract*, scenarios use non-racial, color-based labels such as "advantage purple" or "advantage yellow", see Figures 2 and 14.
- **Candidate Splits:** We define candidate splits as the proportion of candidates across groups. For instance, a 25/75 split means 25% of the candidates are in group 1, while 75% are in group 2. Specifically, we designed a scheme with 8 candidates divided into two groups, forming splits of (*4 vs. 4*: equal-sized groups with 4 candidates each), (*6 vs. 2*: a majority of 6 candidates in one group and a minority of 2 in the other), and (*2 vs. 6*: a minority of 2 candidates and a majority of 6 in the other), see Figures 2, 14.
- **Candidate Performance:** Our scholarship evaluation considers *similarity* in candidates with grades between 90 to 100, focusing on subtle distinctions among top achievers. This approach contrasts with lower grade ranges like 80-90 or 70-80, where such differences are less pronounced. Similarly, *dissimilarity* is examined through a wider grade range of 60 to 100, see Figures 2, 14.

We systematically construct scenarios advantaging one group above another under different grade and fairness constraints. For instance, in "School District K" displayed in *"Table A"* in Figure 1 all white students have higher grades than all black students. This ensures that the purely score-based ranking would be "unfair" if measured by the group-aware fairness metrics ARP or NDKL. We also construct a scenario in which students are ordered for scholarship distribution by descending grades, which concurrently also achieves fairness by both the NDKL and ARP metrics. We refer to this scenario as **"fairness for free"** because students do not need to be re-ranked beyond their grade-based ordering to achieve a metric-assessed fair ranking. This configuration aims to investigate how participants react to fair rankings when the tradeoff between fairness and score-based ordering is removed. This is labeled as Y^* and Z^*, see Figures 2, 14. For both metrics, fairness values range from [0,1] with 0 being ***fair*** and 1 ***not fair***.

Interface & Trial Design

For each split, we generate data for "Table A", "Table B",and "Table C" in Figure 1, including student "names", "grades", "race/group", and "reward". Each table presents two groups of candidates. The first study used a 4 vs. 4 split, resulting in 70 unique combinations. The second and third studies used 6 vs. 2 and 2 vs. 6 splits, respectively, each leading to 28 unique combinations. In total, the study featured 126 unique ranking combinations. The interface design comprised three tables: "students being ranked","proposed ranking of students", and "group-level reward distribution". Specifically:

Students being ranked: "Table A" in Figure 1. This table displays key details about the candidates, including their names, race or group identifiers, and grades in a horizontal format.

Proposed ranking of students: "Table B" in Figure 1 presents candidate rankings based on fairness metrics, with a vertical orientation to indicate order. The table adopts a linear reward order, decreasing by $1,000 per rank from $8,000 for the top candidate to $1,000 for the lowest. (Early pilot studies also investigated a logarithmic reward assignment, but a comparison to linear showed similar patterns with no substantial difference, so we did not include it as a factor in the main study.)

Group-level reward distribution: "Table C" in Figure 1 displays the total scholarship amounts for each group, calculated from the "Table B" reward column. Its inclusion was based on pilot studies indicating that participants occasionally would manually sum these values for their judgment explanations.

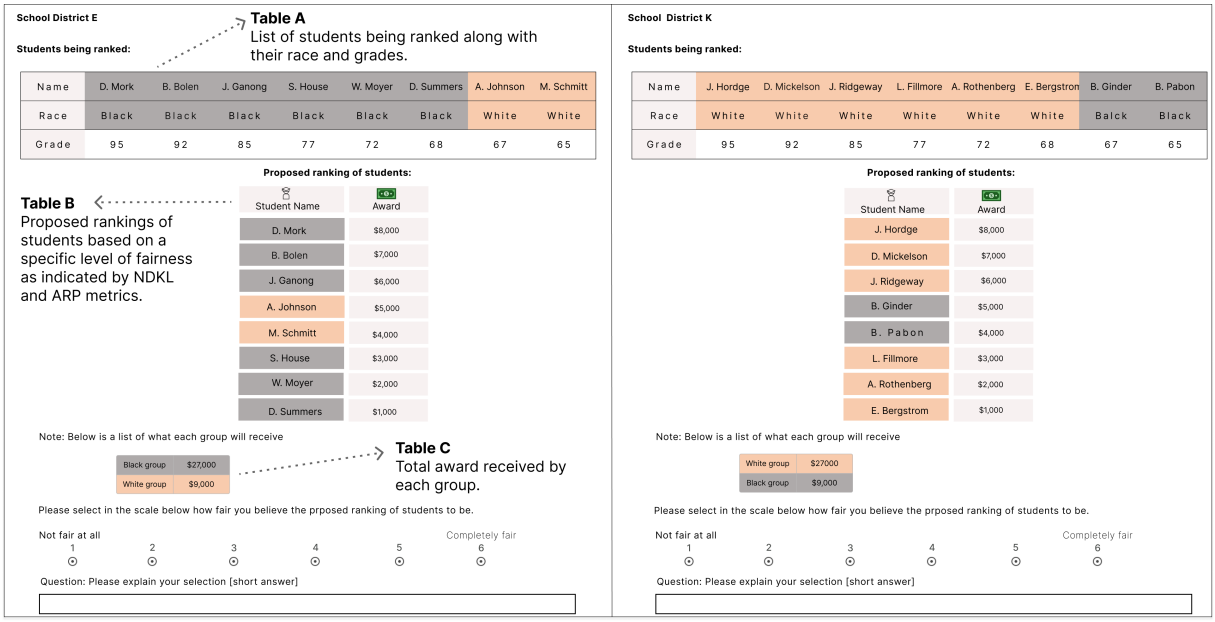

Figure 1: The survey interface, showing a "black advantage" scenario on the left and a "white advantage" scenario on the right.

Open-ended question: Participants were asked to explain their fairness ratings through an optional open-ended question, providing qualitative feedback.

Selection of Metrics

Fairness metrics: We used the two metrics **NDKL** and **ARP** as foundational elements for the construction of the scenarios. Our aim is to investigate rankings in which these metrics agree or disagree with each other, so that we can assess if there are preferences for one of the metrics over the other. We have three distinct cases (using 4 vs. 4 as an example):

- **Both Metrics Yield the Same Value:**
  - *Both Fair*: Both metrics report a value of 0, indicating complete fairness.
  - *Both UnFair*: Each metric indicates a value of 1, suggesting total unfairness.
  - *Fairness for Free*: Both metrics show a value of 0, indicating fairness for both, and the proposed ranking also matches a score-based ranking.
- **Maximum Difference (Δ) Between Metrics:** This case, with one metric at its minimum and the other at its maximum, may influence individuals to express which of these metrics they align with more.
  - *NDKL Fair 1*: NDKL is 0.08, ARP is 0.63.
  - *ARP Fair 1*: ARP is 0.13, NDKL is 0.65.
- **One Metric at 0, the Other Contrasting:** This case provides insight into the complexities and nuances of fairness assessment, where one metric perceives complete fairness (0) while the other disagrees maximally.
  - *NDKL Fair 2*: NDKL is **0,** ARP is 0.25.
  - *ARP Fair 2*: ARP is **0,** NDKL is 0.67.

Table 1: The table displays algorithmic fairness values across three splits: 4 vs. 4, 6 vs. 2, 2 vs. 6, using (ARP) and (NDKL) metrics.

Number	Scenarios	4 vs. 4			6 vs. 2			2 vs. 6		
		Kendall's Tau	ARP	NDKL	Kendall's Tau	ARP	NDKL	Kendall's Tau	ARP	NDKL
1	Both Fair	8	0.00	0.00	5	0.17	0.10	7	0.17	0.10
2	Both Unfair	0	1.00	1.00	12	1.00	1.00	0	1.00	1.00
3	NDKL Fair 1	3	0.63	0.08	11	0.83	0.15	1	0.83	0.15
4	NDKL Fair 2	6	0.25	0.00	0	1.00	0.37	12	1.00	0.37
5	ARP Fair 1	7	0.13	0.65	6	0.00	0.20	6	0.00	0.20
6	ARP Fair 2	8	0.00	0.67	6	0.00	0.20	6	0.00	0.20
7	Fairness for Free	0	0.00	0.00	0	0.17	0.10	0	0.17	0.10

Similar procedures were used to generate rankings for the 6 vs. 2 and 2 vs. 6 conditions, see Table 1. Due to the metrics and imbalances in groups, exact matching values were not possible across all three split conditions.

Kendall's Tau distance metric: Kendall's Tau measures the deviation or difference between two rankings. In our case, we used Kendall's Tau to quantify the difference between our specific ranking scenarios and the score-based ranking in which all candidates are sorted by score alone. A distance of 0 indicates that the two tables are identical, meaning that the order of candidates in *"Table A"* is the same as the order in *"Table B"*. Conversely, a larger distance indicates more disagreement between the score-based ranking and the proposed fair ranking, or put differently, it indicates that a more drastic number of fairness interventions were undertaken to reorder the initial ranking, see Table 1.

Procedure

Instructions: To mitigate outcome favorability bias, we followed [70, 75] by devising a fictional school district scenario. This scenario involved distributing merit scholarships among ranked students,

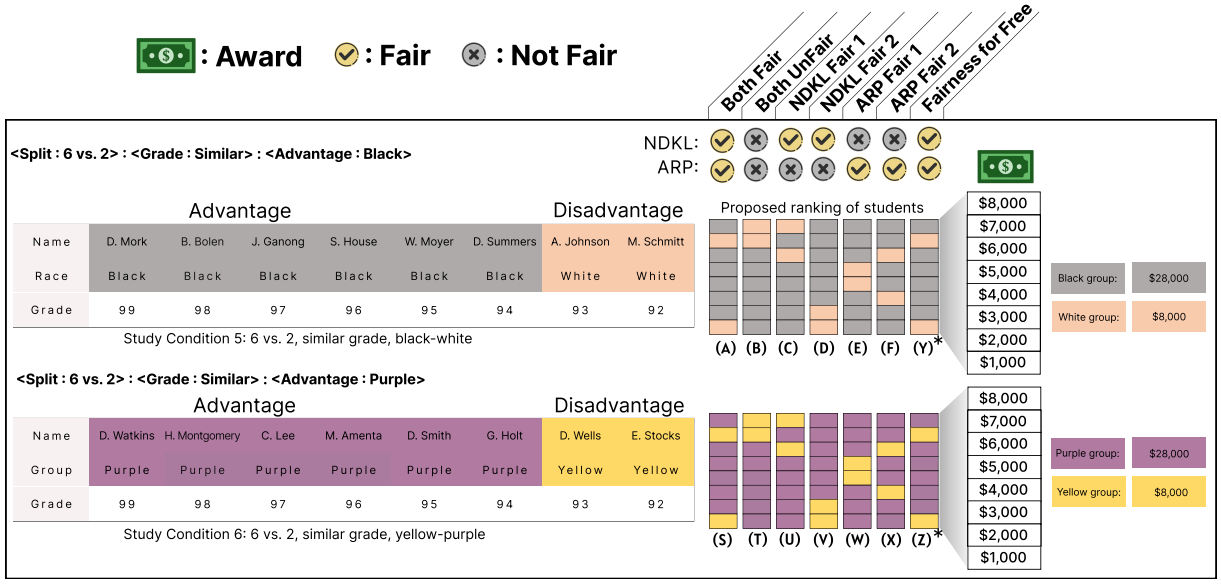

Figure 2: Rankings in a 6 vs. 2 split for both concrete (black vs. white) and abstract (yellow vs. purple) studies, including similar grades. It shows candidate scores and groups on the left, proposed rankings in the center, and award distribution examples on the right. 'Fairness for Free' is indicated by Y* and Z*, representing grade-based rankings. The 2 vs. 6 split mirrors this by inverting rankings.

with variations in split, advantage, and performance conditions, designed to reduce bias in participant responses. Experiment instructions describe the rankings as being determined by a committee (i.e., not by fairness metrics or algorithms).

Structure: The study used a mixed design with 12 conditions, each varying three variables: *split type, grade type,* and *advantage type.* For instance, *condition 1* used a 4 vs. 4 split, similar grades, in a black-white context (Figure 15).

Demographics: After completing the survey, we collected demographic data from participants, including information on gender, age, political views, education, and race, with options for non-disclosure available.

Measures: Participants' fairness perceptions were measured using a six-point Likert scale *(1: 'Not fair at all'* to *6: 'Completely fair')*, known as the 'belief score', serving as the dependent variable for our analysis. Given the known limits of ordinal scales (e.g. [11]), we also include an optional free text response with each question, which participants frequently used to share additional thoughts throughout the study.

Independent Variables: For the Grade variable, there are 'similar' and 'dissimilar' levels. The split variable includes *(4 vs. 4), (6 vs. 2),* and *(2 vs. 6)* levels, and the advantage variable has two levels: *concrete* (e.g., *'advantage black/white'*) and *abstract* (e.g., *'advantage yellow/purple'*).

Recruitment and Participants

Procedure: We conducted the study using the Prolific platform for participant recruitment, with Qualtrics for data collection. The participant pool was individuals who were at least 18 years old, in the U.S., demonstrated high English proficiency with a minimum approval rating of 100. Participants were permitted to participate only once. An IRB-approved informed consent was given, and participant authentication was ensured through the given Prolific registration token and API. Participants were compensated at a rate of $12 per hour, following an estimated median completion time of the study being 20.36±5.43 minutes. Demographics included 239 women, 224 men, 13 non-binary, and 4 other. Age groups spanned 25-34 (176), 18-24 (94), 35-44 (84), 45-54 (66), 55-64 (43), over 64 (16), and 1 undisclosed. Other demographics (e.g. education, race, political affiliation) are included in Appendix D

RESULTS AND ANALYSES

In this section, the analysis approach is outlined, combining quantitative methods and results with a qualitative review of participant comments. The study examined twelve conditions, systematically varying across three key variables: *Split*, *Performance*, and *Advantage*, from Condition 1 to Condition 12. Additionally, we incorporated seven *Scenarios* for each condition, namely: Both Fair, Both Unfair, NDKL Fair 1, NDKL Fair 2, ARP Fair 1, ARP Fair 2, and Fairness for Free. To mitigate order effects, we randomized the order of these scenarios, see Figure 15.

For ***RQ1*** and ***RQ2***, addressing people's perceived notions of fairness and the role of race in perceived fairness, respectively, the analysis utilized a two-way repeated ordinal regression with a Cumulative Link Mixed Model (CLMM). This choice was due to the model's appropriateness for ordinal data [15, 16, 46]. In the model, interaction effects involving the *Advantage* variable were included, hypothesizing that perceptions of fairness might be influenced by the interplay between *Advantage* and other factors. We used *Participants ID* as a random factor to capture diverse individual responses across scenarios. In ***RQ3***, which explored the impact

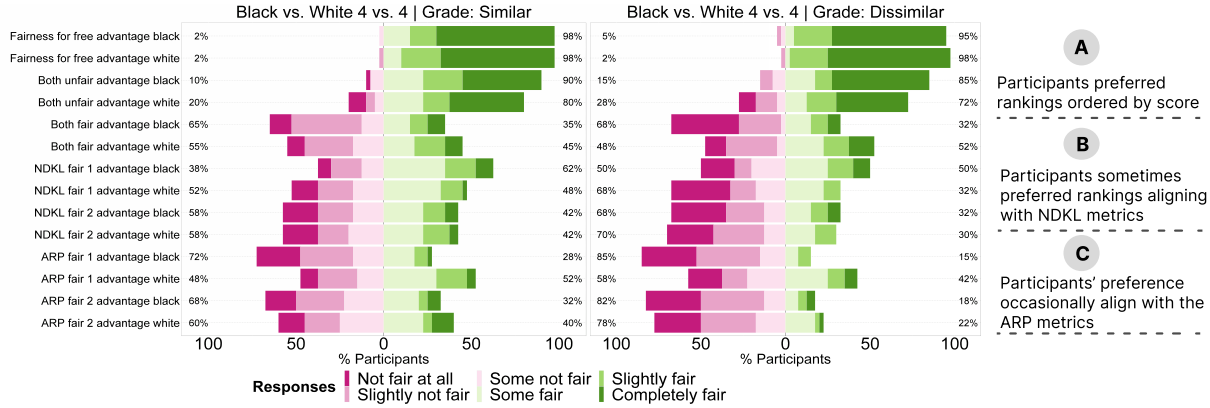

Figure 3: Participant preferences favored rankings based on scores (A), with some support for metrics-based fair rankings (B,C).

of score variability on perceived fairness, we categorized Likert responses into two groups: responses of 4-6 as positive and those lower as negative. We applied the non-parametric Wilcoxon signed-rank test to identify overarching differences in response patterns.

To assess how different variables affect beliefs, we conducted an Analysis of Deviance employing likelihood ratio tests. This method allowed for evaluating the impact of various factors, providing insights into statistical significance and effect sizes. For factors showing significant effects, the "emmeans" library in R was used for post hoc multiple comparisons to explore interactions between factor levels. P-values were adjusted with Tukey Correction. A power analysis was conducted using initial data, aiming for a statistical power of 0.8 and a significance level of 0.05. This analysis indicated that to detect a large effect size ($f^2 = 0.35$), a minimum sample size of $N = 37$ was required for general linear models. To accommodate potential participant dropouts and variations, we recruited 40 participants for each study condition, leading to a total of 480 participants. Those who failed to pass attention checks during the experiment were excluded and replaced, maintaining a final sample size of $n = 480$.

4.1 RQ1: People's Perceived Notions of Fairness:

This condition investigated how perceptions of fairness vary with different fairness metrics, noting that fairness schemes often require reordering candidates. It is expected that individuals will prefer rankings resembling candidates' initial scores. However, the study also explores how different metrics might yield rankings more aligned with participants' preferences, due to varying scores across these metrics. We explore these dimensions across 4 vs. 4, 6 vs. 2 (minority disadvantage) , and 2 vs. 6 (minority advantage) scenarios. This condition is a within-subjects design, as all participants judged rankings across group possibilities, e.g. where race is explicitly mentioned or abstract group labels were used.

Order by Score 4 vs. 4: The two scenarios *"Fairness For Free"* and *"Both Unfair"* were significantly preferred by participants, indicating a belief in their fairness based on order by score, both having a Kendall's Tau value of 0, see Figure 14 for visualizations of the *"Fairness For Free"* scenario, highlighting contrasts Y^* (black vs. white) and Z^* (yellow vs. purple).

Participants consistently preferred the *"Fairness For Free"* scenario across all study conditions, both in *"white vs. black"* in Figure 3 and *"yellow vs. purple"* in Figure 8. *Study condition 4* showed the largest differences (estimate: 3.50, $p < 0.001$), followed by *condition 1* (estimate: 3.32, $p < 0.001$). *Study condition 3* also showed significant differences (estimate: 3.04, $p < 0.001$), as did *study condition 2* (estimate: 2.23, $p < 0.001$); see Table 2.

For the *"Both Unfair"* scenario, we found a consistent perception of fairness among participants across all study conditions. *Study condition 4* showed the largest differences (estimate: 3.31, $p < 0.001$), followed by condition study 1 (estimate: 2.31, $p < 0.001$). *Study condition 3* also showed a significant effect (estimate: 2.29, $p < 0.001$). Lastly, condition study 2 also revealed a notable effect (estimate: 1.04, $p < 0.05$). See Table 2.

In an ANOVA analysis of the CLMM model, fairness beliefs were assessed under various conditions. For *study condition 1*, results indicated a significant effect of 'Scenarios' $\chi^2(6; N = 40) = 276.08$, $p < 0.001$. Similarly, in *study condition 3*, "Scenarios" also showed a significant effect $\chi^2(6; N = 40) = 281.06$, $p < 0.001$. The interaction between "Advantage" and "Scenarios" significantly influenced fairness beliefs in *study condition 1* $\chi^2(6; N = 40) = 13.96$, $p = 0.03$ and in *study condition 3* $\chi^2(6; N = 40) = 24.26$, $p < 0.001$. However, the "Advantage" factor alone did not lead to significant differences in *study condition 1 and 3*. See Tables 3 and 4. In *study condition 2 ,4*, "Scenarios" again demonstrated significant effects, with $\chi^2(6; N = 40) = 147.04$ ($p < 0.001$) in *study condition 2* and $\chi^2(6; N = 40) = 337.65$ ($p < 0.001$) in *study condition 4*. Neither "Advantage" nor "Advantage: Scenario" showed significant differences in these conditions; see Tables 5 and 6.

In a post-hoc analysis, significant differences in fairness perception ratings were noted. In *study condition 1* (ARP vs. NDKL Scenarios), fairness ratings favored NDKL over ARP (estimate = 1.55, $p < 0.01$) Table 7. *Study condition 3* similarly showed a preference for NDKL (estimate = 1.4, $p < 0.03$) Table 8, while *study condition 4* indicated a contrasting trend, with a significant difference favoring ARP (estimate = -1.35, $p < 0.04$) as shown in Table 9.

"Order by Score 6 vs. 2:" The two scenarios *"Fairness For Free"* and *"NDKL fair 2"* were significantly preferred by participants, indicating a belief in their fairness based on order by score, both

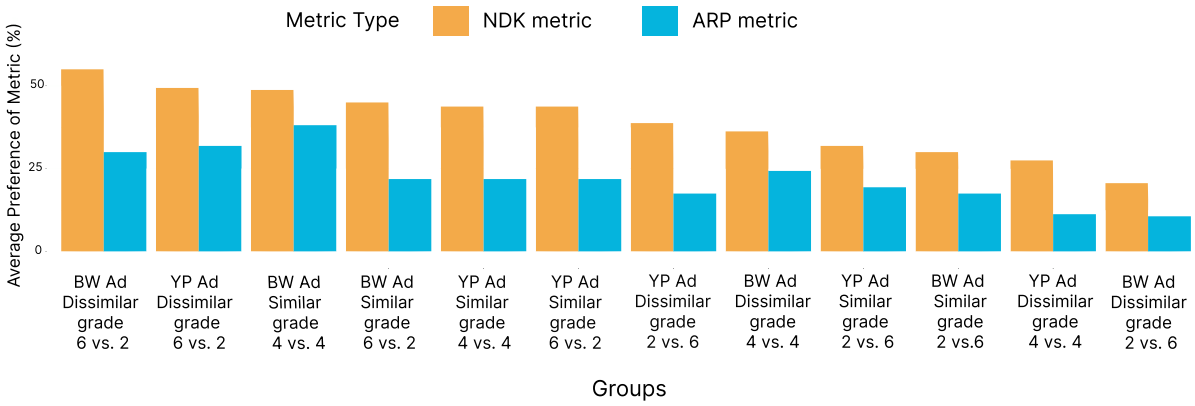

Figure 4: The comparison of NDKL and ARP metrics suggests that NDKL more effectively captures people's perceptions of fairness.

having a Kendall's Tau value of 0, see Figure 2 for visualizations of the *"Fairness For Free"* scenario, highlighting contrasts Y^* (black vs. white) and Z^* (yellow vs. purple). View the contrasts D (black vs. white) and V (yellow vs. purple) in the *"NDKL fair 2"* scenario.

Participants consistently favored the *"Fairness For Free"* scenario in both *"white vs. black"* in Figure 9 and *"yellow vs. purple"* in Figure 10 conditions. The largest differences were observed in *study condition 5* (estimate: 5.45, $p < 0.001$), followed by *study condition 7* (estimate: 5.12, $p < 0.001$). *study condition 8* also showed significant differences (estimate: 3.80, $p < 0.001$), and *study condition 6* showed considerable differences as well (estimate: 3.47, $p < 0.001$). This ordering reflects the varying degrees of perceived fairness across each condition. See Table 10.

In the *"NDKL fair 2"* scenario, participants consistently perceived fairness across all study conditions. Analyses showed that *study condition 7* showed the largest differences (estimate: 5.13, $p < 0.001$), followed by *study condition 5* (estimate: 4.01, $p < 0.001$). *Study condition 8* also showed significant differences (estimate: 3.58, $p < 0.001$), and *study condition 6* also showed notable differences (estimate: 3.51, $p < 0.001$), see Table 10.

In an ANOVA analysis, the results for *study condition 5* indicated a significant effect of "Scenarios" ($\chi^2(6; N = 40) = 331.24$, $p < .001$), suggesting varied perceptions of fairness across scenarios. See Table 11. The analysis also revealed a significant effect of "Advantage" on fairness beliefs ($\chi^2(1; N = 40) = 9.59$, $p < .001$). In *study condition 7*, significant results were observed across multiple parameters: for "Scenarios" ($\chi^2(6; N = 40) = 338.78$, $p < .001$), for "Advantage" ($\chi^2(1; N = 40) = 5.06$, $p = 0.02$), and for the interaction between "Advantage" and "Scenarios" ($\chi^2(6; N = 40) = 38.46$, $p < .001$), see Table 12. However, the interaction between "Advantage" and "Scenarios" was not significant. In study conditions 6 and 8, no significant interactions were found between "Advantage" and "Scenarios", and the effect of "Advantage" itself was also not significant, see also Tables 13 and 14.

A post hoc analysis with correction shows significant differences in fairness perception ratings across study conditions. Conditions (5, 6, 7, 8) are shown in Tables 15, 16, 17, and 18, respectively. Of the 40 pairwise comparisons, 35 indicated a higher fairness perception for NDKL compared to ARP. Only 5 comparisons favored ARP, notably in conditions (5) and (7), as detailed in Tables 15 and 17.

"Order by Score 2 vs. 6:" The two scenarios *"Fairness For Free"* and *"Both unfair"* were significantly preferred by participants, indicating a belief in their fairness based on order by score, both having a Kendall's Tau value of 0, see Figure 2 for visualizations of the *"Fairness For Free"* scenario, highlighting contrasts Y^* (black vs. white) and Z^* (yellow vs. purple).

Analysis reveals consistent preference for the *Fairness For Free* scenario across all study conditions, including *"white vs. black"* (Figure 11) and *"yellow vs. purple"* (Figure 12). The strongest preference was observed in *study condition 9* (estimate: 7.26, $p < 0.001$), followed by *study condition 10* (estimate: 5.94, $p < 0.001$). *study condition 12* also demonstrated a considerable impact (estimate: 4.95, $p < 0.001$), and *study condition 11* showed a notable effect (estimate: 4.26, $p < 0.001$). This ordering reflects the varying degrees of perceived fairness across each condition, see Figure 19.

In the exploration of the *"Both unfair"* scenario, participants consistently perceived fairness across all study conditions. The analysis showed that *study condition 9* had impact (estimate: 6.77, $p < 0.001$), followed by *study condition 10* (estimate: 6.50, $p < 0.001$). *Study condition 11* also revealed significant differences (estimate: 6.55, $p < 0.001$), and *study condition 12* demonstrated notable differences as well (estimate: 5.31, $p < 0.001$). See Table 19.

In the ANOVA analysis, the results for study condition 9 indicated a significant effect of "Scenarios" ($\chi^2(6; N = 40) = 411.89$, $p < .001$), suggesting varied perceptions of fairness across scenarios see Table 20. In *study condition 11* significant results were observed across multiple parameters: for "Scenarios" ($\chi^2(6; N = 40) = 427.34$, $p < .001$), for "Advantage" ($\chi^2(1; N = 40) = 12.32$, $p < .001$), and for the interaction between "Advantage" and "Scenarios" ($\chi^2(6; N = 40) = 23.36$, $p < .001$), see Table 21. In *study conditions 10 and 12* no significant interactions were found between "Advantage" and "Scenarios",and the effect of "Advantage" itself was also not significant. See Tables 22 and 23.

In post hoc analysis with correction, we observed significant differences in fairness perception ratings across study conditions. Of 37 pair comparisons, 28 favored NDKL over ARP in fairness

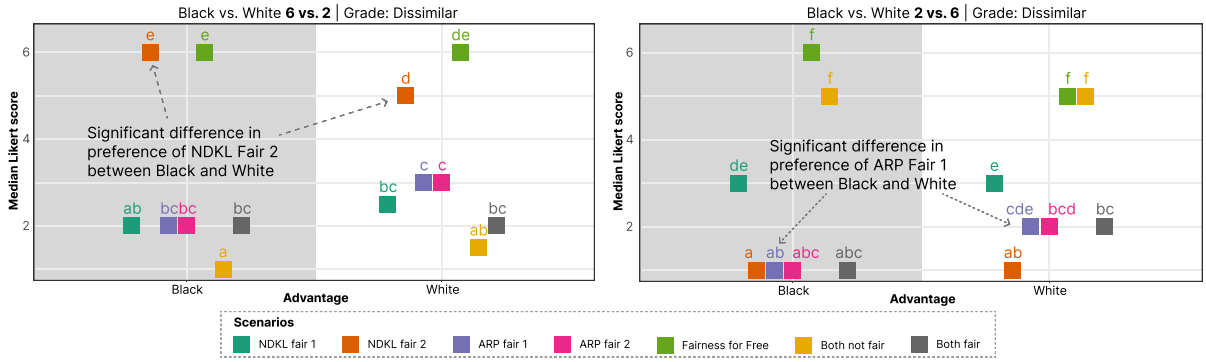

Figure 5: *"NDKL fair 2 scenario"* shows bias towards black candidates, unlike *"ARP fair 1 scenario"* favoring white candidates. Display the data as interaction plots, utilizing group separation letters as referenced in [46, 47]. Differing letters indicate significant score differences between compared scenarios, while shared letters denote no significant differences.

perception (Tables 24, 25, 26, 27), while only 9 comparisons showed a preference for ARP, particularly in *study conditions 9*, *10*, and *11*. See Tables 24, 25, and 26.

"NDKL vs. ARP:" In the assessment of fairness metrics, specifically NDKL and ARP, using the CLMM model and post-hoc analysis, it was found that participants generally preferred rankings generated by the NDKL metric. However, in the 4 vs. 4 split comparisons, the preference for NDKL was not pronounced, indicating fewer statistically significant contrasts. This outcome is attributed to the equal number of candidates in each group, which led participants to favor order by score regardless of "Advantage" group (white, black, yellow, purple). Out of 64 contrasts, 59 supported NDKL, while only 5 supported ARP. Refer to Tables 7, 8, and 9 for detailed results.

In the 6 vs. 2 comparisons, where NDKL was contrasted with ARP across 64 comparisons, participants showed a preference for NDKL 40 times, for ARP 22 times, and 2 instances showed no preference. This suggests that NDKL aligns more with participant perceptions, primarily because it adheres more closely to the order of scores compared to ARP. Similarly, in the 2 vs. 6 comparisons, out of 64 contrasts, participants favored NDKL 36 times and ARP 27 times, with 3 instances showing no distinct preference (Figure 4).

4.2 RQ2: The Role of Race in Perceived Fairness:

Analysis employing a within-subject design focused on scenarios with explicit racial information (e.g., black vs. white) to assess its impact on fairness perceptions. Each participant judged scenarios with alternating racial advantages, particularly in uneven group splits (6 vs. 2 and 2 vs. 6) and scenarios with "dissimilar" grades. We use plots from R's "multcomp" library, with group separation letters for clarity [46, 47]. For example in the "NDKL fair 2" scenario, differing letters denote significant score differences between black and white advantages. Conversely, in the "Fairness for Free" scenario, both groups are denoted by the light green square shape and share the letter 'e,' indicating no significant score difference, see Figure 5 (left).

In the 6 vs. 2 scenario, when black candidates were advantaged, "NKDL fair 2" received a high preference rating of 92%. Conversely, with white candidates advantaged, the preference for "NKDL fair 2" dropped to 70% (*Estimate* = 1.93, *p-value* = 0.004). See Figure 5 (left side). The perceived fairness significantly differs between the orange square shape"e" color (black group) and square shape orange "d" color (white group).

In the 2 vs. 6 split, a different trend was observed. For "ARP fair 1", with black candidates advantaged, the scenario garnered only 5% positive ratings. In contrast, with white candidates advantaged, positive ratings increased to 18% (*Estimate* = -1.62, *p-value* = 0.02), see Figure 5, right side. The contrast is evident between the lilac square shape"ab" (black group) and the lilac square shape"cde" (white group). No other scenarios showed significant results.

4.3 RQ3: Score Variability's Influence on Perceived Fairness: We compared scenarios where candidates had similar scores (ranging between 90-100) against those with dissimilar scores (ranging between 60-100) using between-subjects comparison. This approach involved comparisons such as study condition 1 vs. study condition 3, across a total of 12 groups. The results suggest a significant scaling effect regarding perceived fairness. Specifically, in scenarios where candidates' scores were similar, the rankings were generally perceived as fairer compared to those with more varied scores. This trend was observed both in explicit race-based groups, as evidenced by (Wilcoxon test result of $V = 53, p = 0.01$, Figure 6), and in abstract groups (Figure 7) with a Wilcoxon test result of $V = 55, p = 0.001$. However, in the abstract group scenarios, the pattern was not statistically significant across 6 vs. 2 and 2 vs. 6 conditions, indicating that participants could perceive fairness more significantly with "race" but not with abstract groups. Significant results were observed in the 2 vs. 6 black vs. white scenario when the grade is "similar" (Wilcoxon $V = 50, p = 0.02$, Figure 13).

Qualitative Analysis: During each condition of the experiment, participants had the option to provide a open-response explanation for their choices. These optional responses, totaling 5,981 unique texts after duplicates were removed, shed light on the reasoning behind participants' decisions throughout the experiment. We conducted an open-coding methodology analysis on the data. In this process, we identified several overarching themes, organizing them into two main categories with respective subthemes. We employed a reflexive thematic analysis approach, conducted by three members of the research team, to identify these themes [12, 49]. We

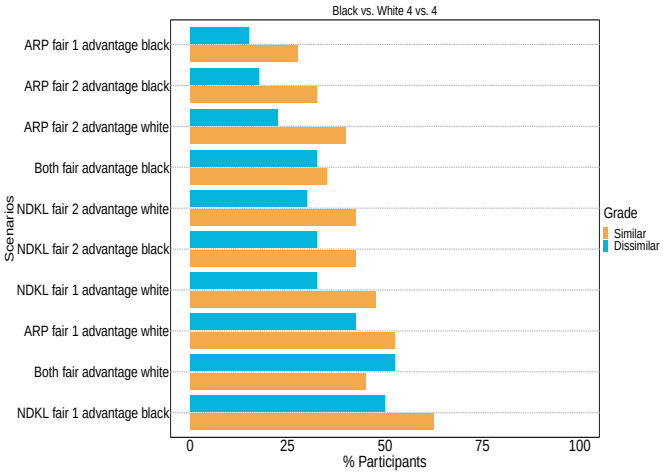

Figure 6: Black vs. White 4 vs. 4 | Grade: Similar and Dissimilar

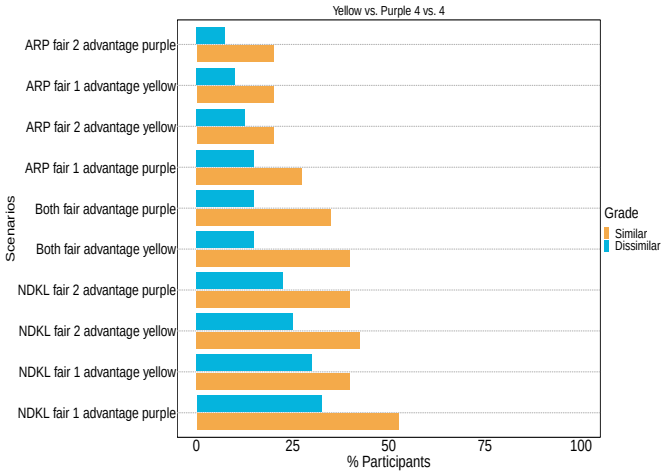

Figure 7: Yellow vs. Purple 4 vs. 4 | Grade: Similar and Dissimilar

assessed the level of agreement among the authors' coding of the responses using the Cohen's kappa statistic. The scores fell between 0.8 and 1, indicating an overall consensus among coders.

Theme 1: Participants believe in meritocracy but find the current ranking system partly unfair due to perceived inconsistencies in score-to-ranking correlations. *First: Merit-based Discontent*: Numerous participants expressed dissatisfaction regarding the perceived absence of merit-based distribution. Participant *P1* disapproved of the method of not awarding funds strictly based on merit. *P3* was surprised to see a student with just a passing score ranked higher than those with mid-high 80s. Similarly, *P6* and *P8* held comparable views about prioritizing candidates with higher grades. See quotes in Table 29. *Second: Perceived Discrimination*: Several participants suggested the presence of bias or discrimination in award distribution. Participant *P2* expressed concerns over potential racism, sexism, or ableism affecting certain groups. Echoing this view, participants *P9* and *P5* shared similar sentiments regarding the fairness of the selection process. *Third: Unequal Distribution*: Several participants identified problems with the unequal distribution among groups. Participant *P4* criticized the allocation of funds to the purple group instead of a fairer distribution to the yellow group. Similarly, Participants *P10* and *P11* expressed concerns about the awards not aligning with performance, questioning the fairness of the allocation process. *Fourth: Arbitrary Ranking*: Participant *P12* points out that the ranking seems arbitrary as *B. Bolen*, despite his grade, is placed at the bottom. This theme revolves around the frustration towards seemingly inconsistent and non-intuitive ranking methods in the award system.

Theme 2: Participants acknowledge the importance of considering socio-economic factors, yet they seek more transparency and fairness in ranking. *First: Inclusion of Multiple Factors for Fair Distribution:* Participants *P16*, *P17*, and *P18* advocate for a comprehensive approach in scholarship distribution, suggesting the inclusion of various factors such as socio-economic status, race, attendance, and challenges associated with minority status to ensure fairness. *Second: Transparency and Adequate Information*: A recurrent theme is the need for transparency and detailed information to comprehend the distribution logic. *P16* calls for clarity on how socioeconomic or minority statuses influence rankings. Likewise, *P22*, *P23*, and *P24* advocate for additional insights into students' socioeconomic backgrounds and educational challenges. *Third: Questioning of Current Ranking Method:* Participants such as *P16* and *P20* expressed doubts about the fairness of the current ranking method. *P16* challenged the assumption that lower grades are tied to class or minority status, and *P20* highlighted an unjust elevation in rankings for certain students.

DISCUSSION: RESULTS, LIMITATIONS AND FUTURE WORK

RQ1: People Largely Prefer to Rank By Score, With Some Exceptions: Across all comparison scenarios, participant perceptions were largely driven by how well each ranking aligned with the underlying scores. Essentially, the further a given ranking was from being ranked by score, the less tolerable participants found it to be. This yields interesting directions worthy for further inquiry. Namely, in the context of our study of students and education, one possibility would be to investigate the extent to which scores themselves could be crafted to better reflect student opportunity and potential, rather than being a raw grade (e.g., standardized test result) alone. Another possibility is to study a more complete record of grades (e.g., across multiple diverse subjects) or grade scales themselves (e.g., would a more unconventional score scale of 1-120 be viewed the same as the more standard scale of (1-100) with the later coming often with a preconceived notion of quality). McConvey et al. [48] reviewed 63 academic papers, showing that algorithms used in higher education heavily rely on student grades and demographic data, with 51% using grades for predicting outcomes and 44% incorporating demographics. This raises concerns about perpetuating historical inequities, a finding supported by our findings.

RQ2: Impact of Race on Perception: This analysis revealed biases in ranking scenarios involving white versus black candidates. Participants perceived scenarios with a 6 vs. 2 split favoring black

candidates as fairer (*scenario D NDKL fair 2"* in Figure 2) than those favoring white candidates. However, a 2 vs. 6 split seen as unfair with a black majority was deemed acceptable for a white majority (*scenario E ARP fair 1"* in Figure 2). This indicates societal biases, perhaps rooted in historical contexts, favoring black candidates. Such asymmetries, statistically significant, reveal that participant judgments are influenced to some degree by factors beyond mere scores. Fugure decision-making system designs might integrate features prompting consideration of fairness and transparency as emphasized by Yurrita et al. [80]. Achieving a balance between system fairness and social realities requires transparent, ethical algorithms that address stakeholders' concerns for providing equitable applications.

RQ3: Candidate Similarity as a Scaling Effect: The analysis indicates scenarios with similar candidate scores are seen as more acceptable than those with diverse scores. This suggests that grouping candidates with similar scores might enhance perceived fairness, informing decision-support system designs. However, the effects of more complex rating systems that mix scores (quantitative data) and letter grades (qualitative data) remain unclear, pointing to a need for further research to understand these complexities.

While these findings contribute to our understanding about the intersection of peoples' perceptions of fairness and fairness metrics, the study is not without limitations, some of which we list here along with ideas for future efforts: For one, our study used a relatively small sample of eight candidates, which may not fully mirror real-world complexities. This highlights the need for studies with larger more diverse samples and a broader range of attributes like gender, religion, and their intersectionality. We used Likert-scale responses which are commonly used to for quantitative comparison, yet there are potential issues with ordinal data and analyses techniques noted in methodology-focused work, e.g. [11, 41, 61]. The development of more sensitive statistical modeling approaches for individual variance, e.g. in Burkner's work [11], may prove useful in future studies. While this study collected demographics (e.g. age, identified race and gender) similar to prior studies [60, 61], our primary focus was on overall effects between metrics and perceptions. More robust models may also enable the more rigorous stratification across demographics of interest.

CONCLUSION

As algorithmic decision-making increasingly impacts people globally, examining the relationship between notions of fairness and people's perceptions is important. Our study assesses how various hypothetical scholarship distribution scenarios, aligning or not with different fairness metrics, are perceived. We found that people tend to perceive scenarios as fairer when they are ordered by score rather than by metrics of fairness, and that factors like demographic information and similarity in candidates' scored performance significantly influence fairness perceptions. Methodologies like ours, linking people's perceptions of fairness to algorithmic fairness metrics, could be pivotal in shaping the development of new algorithmic approaches and consequently in improving decision-support systems.

ACKNOWLEDGMENTS

This work was supported by NSF IIS #2007932, as well as Imam Abdulrahman Bin Faisal University (IAU) and Saudi Arabian Cultural Mission to the USA (SACM).

RESEARCH ETHICS AND SOCIAL IMPACT

Ethical Considerations. We addressed these ethical factors in our study design: *1) Name Selection:* In our study, we utilized a Python library (names) [32] to generate names for survey subject candidates, aiming to minimize racial bias in participant responses. This approach included using last names to mitigate the influence of bias against groups based on race and selecting names that do not give any indication of the candidates' identities. However, this library was not fully equipped to produce names from a comprehensive range of racial backgrounds, such as Hispanic, Black, or White. We recognize this as an opportunity for improvement in research tools. We thus encourage developers interested in fairness and ethical research to develop and enhance libraries capable of generating names that represent a broader spectrum of racial and ethnic backgrounds. *2) Context Selection:* Selecting medium-context scholarships instead of high-stake context aims to prioritize the well-being of participants who have connections to individuals affected by automated decision-making.

Respect for Privacy and Participant Rights: Our study initiated with a detailed consent form that clearly outlined the research objectives and participants' rights, ensuring informed consent. Participation was voluntary for individuals at least 18 years old, with the freedom to withdraw at any time and contact information for the Institutional Review Board (IRB) office provided. Participants were given the option to receive a copy of their consent form, which helped reinforce their understanding and agreement with the study's procedures and goals. During the collection of demographic data, we included a "Prefer not to respond" option for all questions, respecting participant privacy and encouraging unbiased feedback about fairness perceptions. This approach also acknowledged the sensitive nature of demographic data and aimed to build trust among participants. All researchers involved in the study completed CITI Program training to ensure responsible and ethical data handling.

A key aspect of our research methodology was the preservation of participant anonymity. This was achieved by substituting identifiable Prolific IDs with non-identifiable labels (e.g., p1, p2, p3). This practice is in line with our ethical commitment to keep personal and demographic data separate from individual responses, ensuring the protection of participant identities and the integrity of our analysis.

Researcher Positionality. Our team, with expertise in computational solutions, algorithmic design, and Human-Computer Interaction (HCI) centered on human needs, is influenced by the American educational system's focus on equitable group treatment. This shapes our research, particularly in scholarship allocation studies. However, our methodologies and results may not directly translate to international settings due to differing legal and cultural views on fairness and group rewards. Consequently, it is essential

to understand the particular situations of marginalized groups, education systems, and governmental policies in each country before applying our methodologies in these diverse environments.

Adverse Impacts. The use of AI in automated decision-making systems could change how fairness is perceived across different communities. This change might significantly affect how these communities interact with and view these systems. There is a possibility that some groups might feel disadvantaged or misrepresented by the fairness metrics we propose. In response, we advocate for the transparent disclosure of the metrics and methodologies used in these systems. Further, we encourage entities like governments, educational institutions, and relevant committees to have comprehensive discussions about the potential impact of these metrics on marginalized groups before their implementation, ensuring fair and equitable treatment for all.

REFERENCES

[1] Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. 2023. An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (, Chicago, IL, USA,) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1324–1333. https://doi.org/10.1145/3593013.3594083

[2] James D Anderson. 2010. *The education of Blacks in the South, 1860-1935*. Univ of North Carolina Press, Chapel Hill, NC, USA.

[3] Sam Corbett-Davies Avi Feller, Emma Pierson and Sharad Goel. 2016. *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.* Associated Press. Retrieved October 17, 2016 from https://perma.cc/KYC8-KYGD

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[5] Bamzi Banchiri. 2023. *Is Amazon same-day delivery service racist?* The Christian Science Monitor. Retrieved Dec 24, 2023 from https://www.csmonitor.com/Business/2016/0423/Is-Amazon-same-day-delivery-service-racist

[6] Chelsea Barabas. 2020. Beyond Bias: Re-Imagining the Terms of" Ethical AI" in Criminal Law. *Geo. JL & Mod. Critical Race Persp.* 12 (2020), 83.

[7] Teanna Barrett, Quanze Chen, and Amy Zhang. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Chicago</city>, <state>IL</state>, <country>USA</country>, </conf-loc>) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1757–1771. https://doi.org/10.1145/3593013.3594114

[8] Gary S Becker. 2010. *The economics of discrimination*. University of Chicago press, Chicago, IL, USA.

[9] Brooke Bosley, Christina N. Harrington, Susana M. Morris, and Christopher A. Le Dantec. 2022. Healing Justice: A Framework for Collective Healing and Well-Being from Systemic Traumas. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Virtual Event</city>, <country>Australia</country>, </conf-loc>) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 471–484. https://doi.org/10.1145/3532106.3533492

[10] Robin N. Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Chicago</city>, <state>IL</state>, <country>USA</country>, </conf-loc>) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 379–388. https://doi.org/10.1145/3593013.3594005

[11] Paul-Christian Bürkner and Matti Vuorre. 2019. Ordinal regression models in psychology: A tutorial. Advances in Methods and Practices in Psychological Science, 2 (1), 77–101.

[12] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

[13] K. Cachel, E. Rundensteiner, and L. Harrison. 2022. MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, Kuala Lumpur, Malaysia, 1124–1137. https://doi.org/10.1109/ICDE53745.2022.00089

[14] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 390, 17 pages. https://doi.org/10.1145/3411764.3445308

[15] Rune Haubo B Christensen. 2018. Cumulative link models for ordinal regression with the R package ordinal. *Journal of Statistical Software* 35 (2018), 2–45.

[16] Rune Haubo B Christensen. 2019. A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package. *Tutorial for the R Package ordinal https://cran. r-project. org/web/packages/ordinal/Accessed* 1 (2019), 1–10.

[17] Karlijn Dinnissen, Isabella Saccardi, Marloes Vredenborg, and Christine Bauer. 2023. Looking at the FAccTs: Exploring Music Industry Professionals' Perspectives on Music Streaming Services and Recommendations. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter* (<conf-loc>, <city>Athens</city>, <country>Greece</country>, </conf-loc>) *(CHIGREECE '23)*. Association for Computing Machinery, New York, NY, USA, Article 25, 5 pages. https://doi.org/10.1145/3609987.3610011

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[19] Sina Fazelpour and Zachary C. Lipton. 2020. Algorithmic Fairness from a Non-ideal Perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 57–63. https://doi.org/10.1145/3375627.3375828

[20] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[21] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search &; Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &; Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2221–2231. https://doi.org/10.1145/3292500.3330691

[22] Jennifer M Gómez. 2019. What's the harm? Internalized prejudice and cultural betrayal trauma in ethnic minorities. *American Journal of Orthopsychiatry* 89, 2 (2019), 237.

[23] Kishonna L Gray. 2019. Algorithms of oppression: how search engines reinforce racism.

[24] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. https://doi.org/10.1145/3178876.3186138

[25] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*. Neural Information Processing Systems Foundation, Barcelona, Spain, 11.

[26] Leif Hancox-Li and I. Elizabeth Kumar. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 817–826. https://doi.org/10.1145/3442188.3445943

[27] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[28] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. " Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–31.

[29] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[30] James Oliver Horton and Lois E Horton. 2004. *Slavery and the Making of America*. Oxford University Press, New York, NY, USA.

[31] Julie Hui, Jesse King, Cynthia Mcleod, and Amy Gonzales. 2023. High Risk, High Reward: Social Networking Online in Under-resourced Communities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 278, 12 pages. https://doi.org/10.1145/3544548.3581084

[32] Trey Hunner. 2024. *names 0.3.0*. Python Package Index (PyPI). Retrieved Jan 6, 2024 from https://pypi.org/project/names/

[33] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of*

the ACM on Human-Computer Interaction 2, CSCW (2018), 1–18.

[34] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[35] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[36] Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 674–682. https://doi.org/10.1145/3461702.3462588

[37] Caitlin Kuhlman and Elke Rundensteiner. 2020. Rank aggregation algorithms for fair consensus. *Proc. VLDB Endow.* 13, 12 (jul 2020), 2706–2719. https://doi.org/10.14778/3407790.3407855

[38] Peter Kuhn. 1987. Sex discrimination in labor markets: The role of statistical evidence. *The American Economic Review* 77, 4 (1987), 567–583.

[39] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems* 30 (2017), 1–11.

[40] Michael LaCour-Little. 1999. Discrimination in mortgage lending: A critical review of the literature. *Journal of Real Estate Literature* 7, 1 (1999), 15–49.

[41] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[42] U.S. Federal Legislation. 2022. *U.S. Federal Legislation.* U.S. Department of Justice. Retrieved Dec 1, 2023 from http://www.usdoj.gov

[43] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[44] Malay K Majmundar and David Weisburd. 2018. *Proactive policing: Effects on crime and communities.* National Academies Press, Washington, D.C.

[45] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management* 58, 5 (2021), 102642.

[46] Salvatore S. Mangiafico. 2016. *Two-way Repeated Ordinal Regression with CLMM.* R Companion. Retrieved March 6, 2023 from https://rcompanion.org/handbook/G_12.html

[47] S Mangiafico Salvatore. 2016. Summary and Analysis of Extension Program Evaluation in R.

[48] Kelly McConvey, Shion Guha, and Anastasia Kuzminykh. 2023. A Human-Centered Review of Algorithms in Decision-Making in Higher Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 223, 15 pages. https://doi.org/10.1145/3544548.3580658

[49] Hani Morgan. 2022. Understanding thematic analysis and the debates involving its use. *The Qualitative Report* 27, 10 (2022), 2079–2090.

[50] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New Orleans, USA, 1–17.

[51] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. Association for the Advancement of Artificial Intelligence, New York, USA, 5248–5255.

[52] S Rebecca Neusteter and Megan O'Toole. 2019. Every Three Seconds: Unlocking Police Data on Arrests. Vera Institute of Justice. Available at: https://www.vera.org/publications/arresttrends-every-three-seconds-landing/arrest-trends-every-three-seconds/overview.

[53] Safiya Umoja Noble. 2018. *Algorithms of Oppression.* New York University Press, New York, USA. https://doi.org/10.18574/nyu/9781479833641.001.0001

[54] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[55] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. *Measuring Discrimination in Socially-Sensitive Decision Records.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 581–592. https://doi.org/10.1137/1.9781611972795.50 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611972795.50

[56] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[57] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature human behaviour* 4, 7 (2020), 736–745.

[58] Amifa Raj and Michael D. Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 726–736. https://doi.org/10.1145/3477495.3532018

[59] MIT Technology Review. 2024. *Facebook's ad-serving algorithm discriminates by gender and race.* MIT Technology Review. Retrieved Dec 1, 2023 from https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/

[60] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, PMLR, Online, 8377–8387.

[61] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/3306618.3314248

[62] BBC News Services. 2023. *Google apologises for Photos app's racist blunder.* BBC News Services. Retrieved Dec 10, 2023 from https://www.bbc.com/news/technology-33347866

[63] Vishal Sharma, Kirsten E Bray, Neha Kumar, and Rebecca E Grinter. 2022. Romancing the algorithm: Navigating constantly, frequently, and silently changing algorithms for digital work. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[64] Ellen Simpson and Bryan Semaan. 2021. For you, or for" you"? Everyday LGBTQ+ encounters with TikTok. *Proceedings of the ACM on human-computer interaction* 4, CSCW3 (2021), 1–34.

[65] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[66] Jessie J. Smith, Lex Beattie, and Henriette Cramer. 2023. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *Proceedings of the ACM Web Conference 2023* (<conf-loc>, <city>Austin</city>, <state>TX</state>, <country>USA</country>, </conf-loc>) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3648–3659. https://doi.org/10.1145/3543507.3583204

[67] Jessie J. Smith, Anas Buhayh, Anushka Kathait, Pradeep Ragothaman, Nicholas Mattei, Robin Burke, and Amy Voida. 2023. The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Chicago</city>, <state>IL</state>, <country>USA</country>, </conf-loc>) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1652–1663. https://doi.org/10.1145/3593013.3594106

[68] Nasim Sonboli, Jessie J. Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and Transparency in Recommendation: The Users' Perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 274–279. https://doi.org/10.1145/3450613.3456835

[69] Gregory D Squires. 2003. Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs* 25, 4 (2003), 391–410.

[70] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[71] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

[72] Michael A Stoll, Steven Raphael, and Harry J Holzer. 2004. Black job applicants and the hiring officer's race. *ILR Review* 57, 2 (2004), 267–287.

[73] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden) *(FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[74] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. 2020. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New*

England Journal of Medicine 383, 9 (2020), 874–882. https://doi.org/10.1056/NEJMms2004740 arXiv:https://www.nejm.org/doi/pdf/10.1056/NEJMms2004740

[75] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[76] Isabel Wilkerson. 2010. The warmth of other suns: the epic story of America's great migration. *African Diaspora Archaeology Newsletter* 13, 4 (2010), 25.

[77] Pak-Hang Wong. 2020. Democratizing algorithmic fairness. *Philosophy & Technology* 33 (2020), 225–244.

[78] Carter G. Woodson. 1933. *The Mis-Education of the Negro.* The Associated Publishers, Washington, D.C. 108 pages.

[79] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526

[80] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. https://doi.org/10.1145/3544548.3581161

[81] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938

[82] Marilyn Zhang. 2022. Affirmative Algorithms: Relational Equality as Algorithmic Fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Seoul</city>, <country>Republic of Korea</country>, </conf-loc>) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 495–507. https://doi.org/10.1145/3531146.3533115

[83] Jianlong Zhou, Sunny Verma, Mudit Mittal, and Fang Chen. 2021. Understanding relations between perception of fairness and trust in algorithmic decision making. In *2021 8th International Conference on Behavioral and Social Computing (BESC)* (29-31 October 2021). IEEE, IEEE, Doha, Qatar, 1–5. https://doi.org/10.1109/BESC53957.2021.9635182

[84] Annette Zimmermann and Chad Lee-Stronach. 2022. Proceed with caution. *Canadian Journal of Philosophy* 52, 1 (2022), 6–25.

A APPENDIX 1

Here, we present the specific metric formulations we employ in our study; the Normalized Discounted Cumulative KL-divergence (NDKL) [21] and Attribute Rank Parity (ARP) [13] metrics.

Notation. The metrics consider a finite candidate set $X = x_1, x_2, ..., x_n$ ordered in a ranked list τ. $\tau(x_i)$ denotes the ordinal position of candidate x_i in the ranking τ. Candidates, each belong to a group defined by a shared protected attribute value, such as (gender = "woman") or (gender = "woman" and race = "Asian"). We represent the set of groups associated with the candidates as $G = g_1, g_2, ..., g_m$. We use $D_X = (p_{X:g_1}, ..., p_{X:g_m})$ to represent the distribution of groups in candidate set X, where the proportion of each group is $p_{X:g_m} = |g_m|/|X|$. For example, $D_X = (0.2, 0.3, 0.5)$ indicates that g_1 is 20% of X, and g_2 and g_3 are 30% and 50% of X, respectively.

Metric Formulations. NDKL assesses the representation of different groups at every prefix of the ranking, weighting the higher-up prefixes more. It deems a ranking fair if each prefix, i.e, a top-k set, of the ranking has a proportional share of all groups. NDKL is most fair at 0 and conceptually focused on representing groups fairly higher up in the ranking.

- **NDKL** [21]: of ranking τ with respect to groups G is defined as:

$$NDKL(\tau, G) = \frac{1}{Z} \sum_{i=1}^{|X|} \frac{1}{log_2(i+1)} d_{KL}(D_{\tau_i} || D_X) \tag{1}$$

where $d_{KL}(D_{\tau_i} || D_X)$ is the KL-divergence score of the group proportions of the first i positions in τ and the group proportions of the item set X and $Z = \sum_{i=1}^{|\tau|} \frac{1}{log_2(i+1)}$. Higher $NDKL$ values indicate larger disparities between how groups are represented at top positions compared to their overall size.

In contrast, ARP is a pairwise metric, that decomposes the ranking into pairwise comparisons. It measures the difference between average mixed pairs won by each group, where a mixed pair compares candidates of disjoint groups. It is most fair at 0 and conceptually geared toward ensuring fair group treatment equally across the entire ranking.

- **ARP** [13]: is the maximum absolute difference between average mixed pairs won by each group, calculated as $avgpairs(\tau, g_i) = \#\ mixed\_pairs\_won(g_i) / \#\ total\_mixed\_pairs(g_i)$. Then ARP of ranking τ for groups G is:

$$ARP(\tau, G) = \underset{\forall\ g_j, g_k \in G}{\operatorname{argmax}} |avgpairs(\tau, g_j) - avgpairs(\tau, g_k)| \tag{2}$$

Higher ARP values, up to 1, indicate one or more groups are concentrated at the bottom of the ranking.

B APPENDIX 2

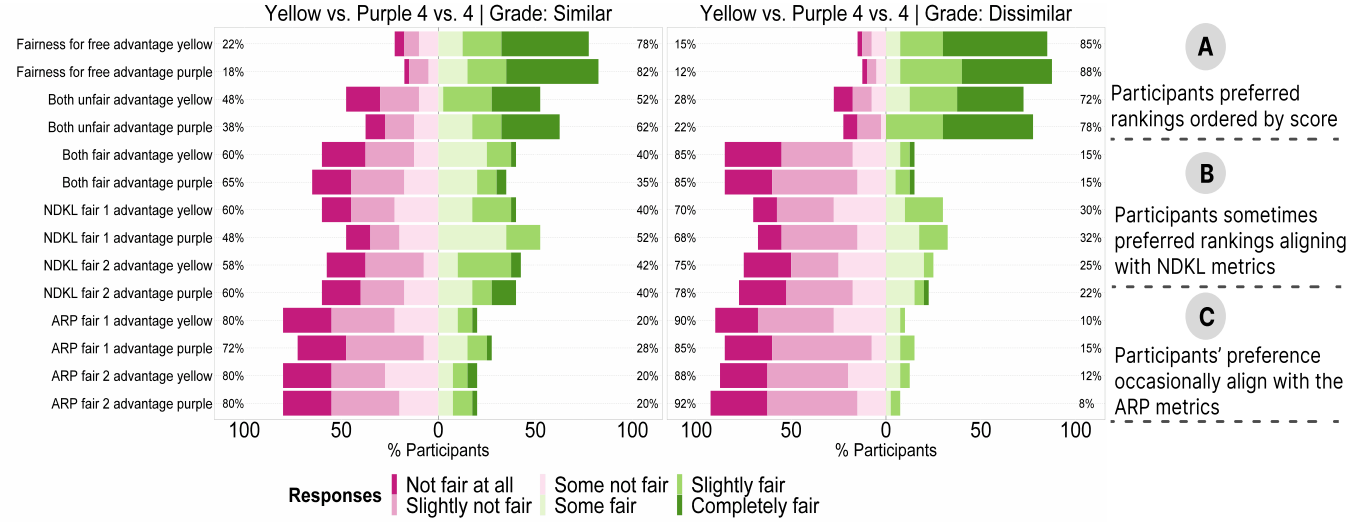

Figure 8: Participant preferences favored rankings based on scores (A), with some support for algorithmically fair rankings (B, C).

Table 2: Coefficient table of the CLMM model in 4 vs.4

Condition	Estimate	Std. Error	Z value	Pr(>|Z|)
Study Condition<1>: Split<4 vs. 4>,Grade<similar grade>,Advantage<black-white>				
Scenarios				
Fairness for free	3.32	6.98	0.48	2.90e-12 ***
Both unfair	2.31	0.45	5.15	2.66e-07 ***
Scenarios: Advantage				
Fairness for free: White	0.61	0.65	0.95	0.34
Both unfair: White	0.15	0.23	0.63	0.82
Random effects (participant id)				
Variance	1.79		Std.Dev	1.34
Study Condition<2>: Split<4 vs. 4>,Grade<similar grade>,Advantage<yellow-purple>				
Scenarios				
Fairness for free	2.23	0.42	5.31	1.1e-07 ***
Both unfair	1.04	0.42	2.47	0.01345 *
Scenarios: Advantage				
Fairness for free: Yellow	0.13	0.59	0.22	0.83
Both unfair: Yellow	-0.21	0.60	-0.35	0.73
Random effects (participant id)				
Variance	1.12		Std.Dev	1.06
Study Condition<3>: Split<4 vs. 4>,Grade<dissimilar grade>,Advantage<black-white>				
Scenarios				
Fairness for free	3.04	0.45	6.68	2.34e-11 ***
Both unfair	2.29	0.44	5.20	2.03e-07 ***
Scenarios: Advantage				
Fairness for free: White	1.20	0.64	1.89	0.06 .
Both unfair: White	0.14	0.61	0.23	0.82
Random effects (participant id)				
Variance	0.44		Std.Dev	0.66
Study Condition<4>: Split<4 vs. 4>,Grade<dissimilar grade>,Advantage<yellow-purple>				
Scenarios				
Fairness for free	3.50	0.44	7.89	3.05e-15 ***
Both unfair	3.31	0.45	7.32	2.47e-13 ***
Scenarios: Advantage				
Fairness for free: Yellow	0.11	0.60	0.19	0.85
Both unfair: Yellow	-0.94	0.60	-1.56	0.12
Random effects (participant id)				
Variance	0.67		Std.Dev	0.82

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 3: Analysis of Condition 1: Split <4 vs. 4> similar grade black-white

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	1.24	1	0.27
Scenarios	276.08	6	< 2e-16 ***
Advantage: Scenarios	13.96	6	0.03 *

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 4: Analysis of Condition 3: Split <4 vs. 4> dissimilar grade black-white

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	1.08	1	0.30
Scenarios	281.06	6	< 2.2e-16 ***
Advantage: Scenarios	24.26	6	0.00047 ***
Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1			

Table 5: Analysis of study condition 2: Split <4 vs. 4> similar grade yellow-purple

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	0.28	1	0.60
Scenarios	147.04	6	<2e-16 ***
Advantage: Scenarios	1.01	6	0.99
Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1			

Table 6: Analysis of study condition 4: <4 vs. 4> dissimilar grade yellow-purple

Facttor	LR Chisq	Df	Pr(>Chisq)
Advantage	0.05	1	0.83
Scenarios	337.65	6	<2e-16 ***
Advantage: Scenarios	4.63	6	0.59
Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1			

Table 7: Study Condition<1>: Split<4 vs. 4>,Grade<similar grade>,Advantage<black-white>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 1 Black - ARP fair 1 Black	1.5460	0.4050	3.8169	0.0101	62.5%	27.5%

Table 8: Study Condition<3>: Split<4 vs. 4>,Grade<dissimilar grade>,Advantage<black-white>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 1 Black - ARP fair 1 Black	1.3846	0.3945	3.5101	0.0301	50%	15%

Table 9: Study Condition<4>: Split<4 vs. 4>,Grade<dissimilar grade>,Advantage<yellow-purple>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
ARP fair 2 purple - NDKL fair 1 yellow	-1.35	0.39	-3.43	0.04	30%	7.5%

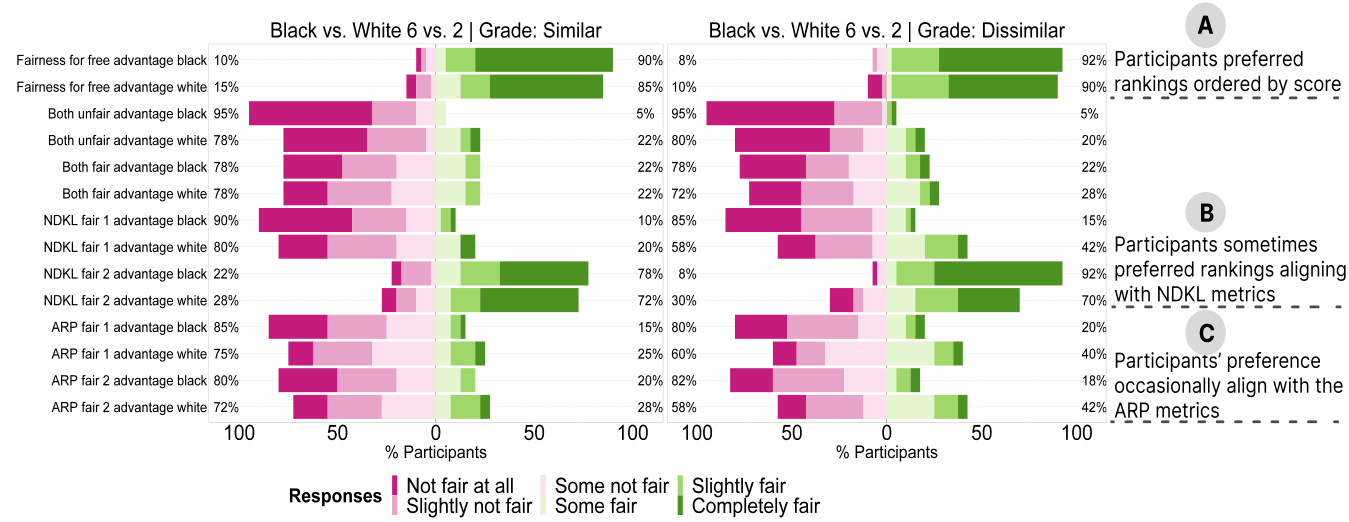

Figure 9: Participant preferences favored rankings based on scores (A,B), with some support for algorithmically fair rankings (C).

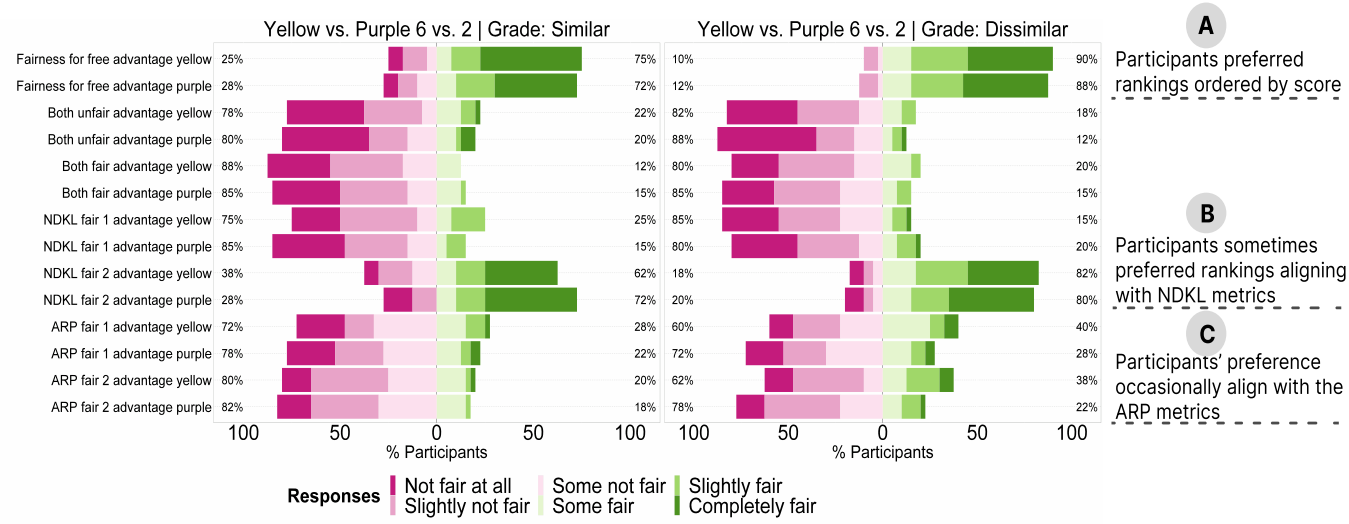

Figure 10: Participant preferences favored rankings based on scores (A,B), with some support for algorithmically fair rankings (C).

Table 10: Coefficient table of the CLMM model in 6 vs.2

Condition	Estimate	Std. Error	Z value	Pr(>|Z|)
Study Condition<5>: Split<6 vs. 2>,Grade<similar grade>,Advantage<black-white>				
Scenarios				
Fairness for free	5.45	0.52	10.39	<2e-16 ***
NDKL fair 2	4.01	0.47	8.57	<2e-16 ***
Scenarios : Advantage				
Fairness for free: White	-1.59	0.65	-2.43	0.0150 *
NDKL fair 2: White	-0.72	0.62	-1.16	0.25
Random effects (participant id)				
Variance	0.60		Std.Dev	0.78
Study Condition<6>: Split<6 vs. 2>,Grade<similar grade>,Advantage<yellow-purple>				
Scenarios				
Fairness for free	3.47	0.45	7.69	1.51e-14 ***
NDKL fair 2	3.51	0.47	7.46	8.76e-14 ***
Scenarios : Advantage				
Fairness for free: Yellow	-0.22	0.61	-0.36	0.72
NDKL fair 2: Yellow	-1.13	0.62	-1.81	0.0699 .
Random effects (participant id)				
Variance	1.23		Std.Dev	1.11
Study Condition<7>: Split<6 vs. 2>,Grade<dissimilar grade>,Advantage<black-white>				
Scenarios				
Fairness for free	5.12	0.50	10.27	<2e-16 ***
NDKL fair 2	5.13	0.50	10.19	<2e-16 ***
Scenarios : Advantage				
Fairness for free: White	-1.74	0.64	-2.74	0.00620 **
NDKL fair 2: White	5.13	0.50	10.19	5.36e-07 ***
Random effects (participant id)				
Variance	1.05		Std.Dev	1.03
Study Condition<8>: Split<6 vs. 2>,Grade<dissimilar grade>,Advantage<yellow-purple>				
Scenarios				
Fairness for free	3.80	0.45	8.47	<2e-16 ***
NDKL fair 2	3.58	0.46	7.76	8.51e-15 ***
Scenarios : Advantage				
Fairness for free: Yellow	-0.03	0.59	-0.05	0.96
NDKL fair 2: Yellow	-0.30	0.60	-0.50	0.62
Random effects (participant id)				
Variance	0.56		Std.Dev	0.75

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 11: Analysis of condition 5: Split <6 vs. 2> similar grade black-white

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	9.59	1	0.00196 **
Scenarios	331.24	6	< 2.2e-16 ***
Advantage: Scenarios	10.05	6	0.12

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 12: Analysis of Condition 7: <6 vs. 2> dissimilar grade black-white

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	5.06	1	0.024 *
Scenarios	338.78	6	< 2.2e-16 ***
Advantage: Scenarios	38.46	6	9.119e-07 ***
Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1			

Table 15: Study Condition<5>: Split<6 vs. 2>, Grade<similar grade>, Advantage<black-white>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 1 Black - ARP fair 2 White	-1.479	0.413	-3.583	0.02347488320	10%	27.5%
NDKL fair 1 Black - ARP fair 1 White	-1.500	0.410	-3.663	0.01772745072	10%	25%
NDKL fair 2 Black - ARP fair 1 White	2.510	0.431	5.821	0.00000052802	77.5%	25%
NDKL fair 2 Black - ARP fair 2 White	2.532	0.433	5.844	0.00000045999	77.5%	27.5%
NDKL fair 2 White - ARP fair 1 White	2.675	0.453	5.909	0.00000031226	72.5%	25%
NDKL fair 2 White - ARP fair 2 White	2.697	0.455	5.932	0.00000027046	72.5%	27.5%
NDKL fair 2 Black - ARP fair 2 Black	3.208	0.445	7.214	0.00000000005	77.5%	20%
ARP fair 2 Black - NDKL fair 2 White	-3.373	0.467	-7.226	0.00000000005	72.5%	20%
ARP fair 1 Black - NDKL fair 2 White	-3.471	0.471	-7.376	0.00000000001	72.5%	15%
NDKL fair 2 Black - ARP fair 1 Black	3.305	0.448	7.381	0.00000000001	77.5%	15%

Table 16: Study Condition<6>: Split<6 vs. 2>,Grade<similar grade>,Advantage<yellow-purple>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 2 yellow - ARP fair 1 yellow	1.977	0.432	4.574	0.00040760648	62.5%	27.5%
ARP fair 1 purple - NDKL fair 2 yellow	-2.188	0.435	-5.030	0.00004334704	62.5%	22.5%
NDKL fair 2 yellow - ARP fair 2 yellow	2.152	0.426	5.049	0.00003936976	62.5%	20%
ARP fair 2 purple - NDKL fair 2 yellow	-2.251	0.426	-5.291	0.00001088854	62.5%	17.5%
NDKL fair 2 purple - ARP fair 1 yellow	2.515	0.445	5.646	0.00000147889	72.5%	27.5%
NDKL fair 2 purple - ARP fair 1 purple	2.727	0.449	6.073	0.00000011351	72.5%	22.5%
NDKL fair 2 purple - ARP fair 2 yellow	2.691	0.441	6.102	0.00000009478	72.5%	20%
NDKL fair 2 purple - ARP fair 2 purple	2.790	0.440	6.336	0.00000002145	72.5%	17.5%

Table 13: Analysis of study condition 6: Split <6 vs. 2> similar grade yellow-purple

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	0.59	1	0.44
Scenarios	224.01	6	<2e-16 ***
Advantage: Scenarios	3.84	6	0.70
Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1			

Table 14: Analysis of study condition 8: <6 vs. 2> dissimilar grade yellow-purple

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	1.68	1	0.20
Scenarios	289.68	6	<2e-16 ***
Advantage: Scenarios	2.51	6	0.87
Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1			

Table 17: Study Condition<7>: Split<6 vs. 2>,Grade<dissimilar grade>,Advantage<black-white>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 1 Black - ARP fair 2 White	-1.420	0.408	-3.479	0.03330458376	15%	42.5%
NDKL fair 2 White - ARP fair 1 White	1.573	0.419	3.757	0.01258406586	7%0	40%
NDKL fair 1 Black - ARP fair 1 White	-1.627	0.403	-4.039	0.00421437219	15%	40%
NDKL fair 2 White - ARP fair 2 White	1.780	0.426	4.175	0.00240940515	70%	42.5%
ARP fair 2 Black - NDKL fair 2 White	-2.433	0.429	-5.666	0.00000131636	70%	17.5%
ARP fair 1 Black - NDKL fair 2 White	-2.629	0.436	-6.028	0.00000015049	70%	20%
NDKL fair 2 Black - ARP fair 1 White	3.502	0.464	7.548	0.00000000000	92.5%	40%
NDKL fair 2 Black - ARP fair 2 White	3.709	0.472	7.849	0.00000000000	92.5%	42.5%
NDKL fair 2 Black - ARP fair 2 Black	4.361	0.478	9.128	0.00000000000	92.5%	17.5%
NDKL fair 2 Black - ARP fair 1 Black	4.558	0.485	9.400	0.00000000000	92.5%	20%

Table 18: Study Condition<8>: Split<6 vs. 2>,Grade<dissimilar grade>,Advantage<yellow-purple>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 2 yellow - ARP fair 1 yellow	2.199	0.418	5.255	0.00001323242	82.5%	40%
NDKL fair 2 yellow - ARP fair 2 yellow	2.347	0.430	5.458	0.00000431626	82.5%	37.5%
NDKL fair 2 purple - ARP fair 1 yellow	2.414	0.435	5.545	0.00000264439	80%	40%
NDKL fair 2 purple - ARP fair 2 yellow	2.561	0.446	5.741	0.00000085191	80%	37.5%
ARP fair 1 purple - NDKL fair 2 yellow	-2.576	0.424	-6.077	0.00000011097	82.5%	27.5%
NDKL fair 2 purple - ARP fair 1 purple	2.791	0.440	6.337	0.00000002127	80%	27.5%
ARP fair 2 purple - NDKL fair 2 yellow	-2.748	0.423	-6.500	0.00000000729	82.5%	22.5%
NDKL fair 2 purple - ARP fair 2 purple	2.963	0.440	6.736	0.00000000148	80%	22.5%

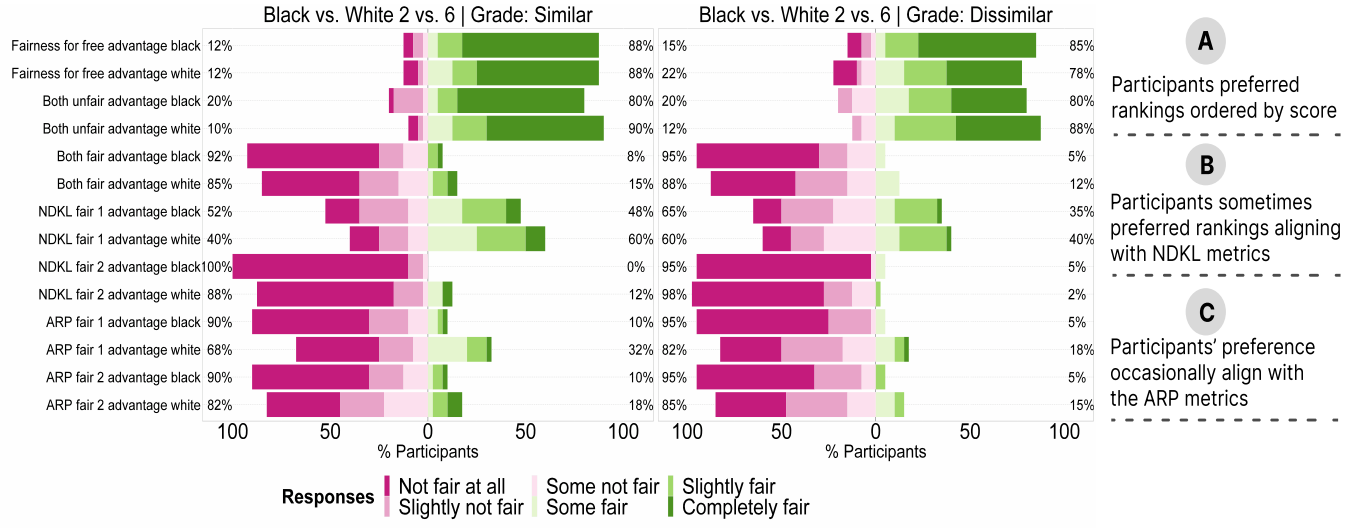

Figure 11: Participant preferences favored rankings based on scores (A,B), with some support for algorithmically fair rankings (C).

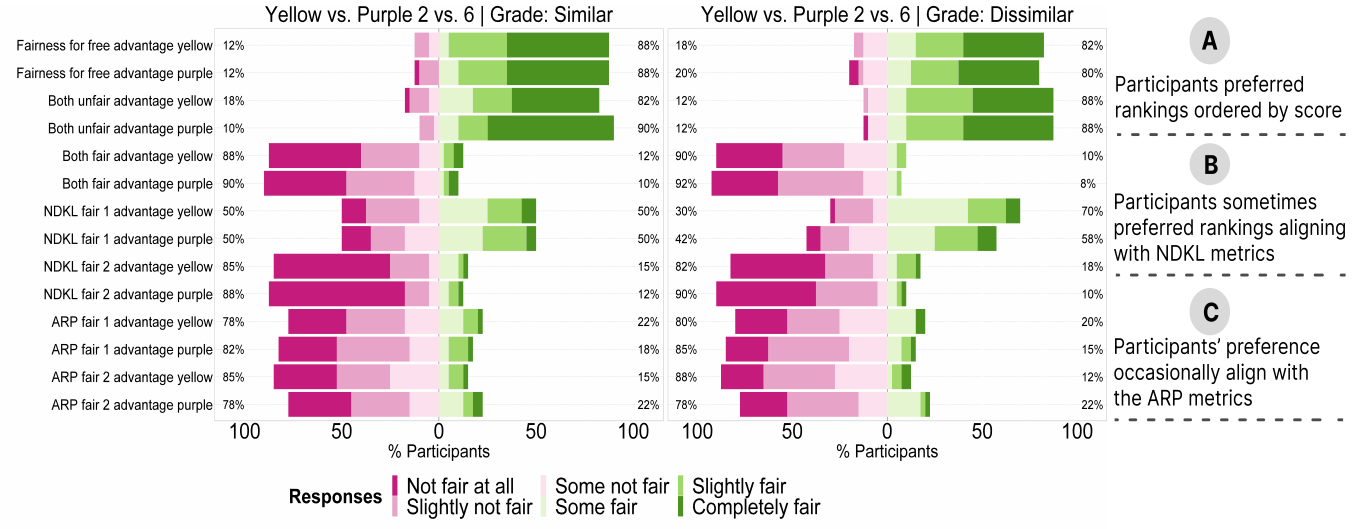

Figure 12: Participant preferences favored rankings based on scores (A,B), with some support for algorithmically fair rankings (C).

Table 20: Analysis of Condition 9: Split <2 vs. 6> similar grade black-white

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	11.72	1	0.0006 ***
Scenarios	411.89	6	< 2.2e-16 ***
Advantage: Scenarios	10.81	6	0.09 .

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 19: Coefficient table of the CLMM model in 2 vs.6

Condition	Estimate	Std. Error	Z value	Pr(>|Z|)
Study Condition<9>: Split<2 vs. 6>,Grade<similar grade>,Advantage<black-white>				
Scenarios				
Fairness for free	7.26	0.73	9.99	< 2e-16 ***
Both unfair	6.78	0.71	9.52	< 2e-16 ***
Scenarios : Advantage				
Fairness for free: White	-2.19	0.84	-2.61	0.009 **
Both unfair: White	-1.7302	0.8283	-2.089	0.04 *
Random effects (participant id)				
Variance	1.23		Std.Dev	1.11
Study Condition<10>: Split<2 vs. 6>,Grade<similar grade>,Advantage<yellow-purple>				
Scenarios				
Fairness for free	5.940	0.544	10.91	< 2e-16 ***
Both unfair	6.50	0.57	11.48	< 2e-16 ***
Scenarios : Advantage				
Fairness for free: Yellow	-0.21	0.68	-0.31	0.75
Both unfair: Yellow	-0.48	0.63	-0.76	0.45
Random effects (participant id)				
Variance	2.12		Std.Dev	1.46
Study Condition<11>: Split<2 vs. 6>,Grade<dissimilar grade>,Advantage<black-white>				
Scenarios				
Fairness for free	4.27	0.68	6.25	3.99e-10 ***
Both unfair	6.55	0.71	9.18	< 2e-16 ***
Scenarios : Advantage				
Fairness for free: White	-2.75	0.85	-3.26	0.00112 **
Both unfair: White	-1.23	0.82	-1.49	0.14
Random effects (participant id)				
Variance	0.62		Std.Dev	0.79
Study Condition<12>: Split<2 vs. 6>,Grade<dissimilar grade>,Advantage<yellow-purple>				
Scenarios				
Fairness for free	4.95	0.48	10.29	< 2e-16 ***
Both unfair	5.31	0.48	10.97	< 2e-16 ***
Scenarios : Advantage				
Fairness for free: Yellow	-0.36	0.62	-0.58	0.56
Both unfair: Yellow	-0.52	0.62	-0.84	0.40
Random effects (participant id)				
Variance	0.64		Std.Dev	0.80

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 21: Analysis of Condition 11: <2 vs. 6> dissimilar grade black-white

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	12.32	1	0.0005 ***
Scenarios	427.34	6	< 2.2e-16 ***
Advantage: Scenarios	23.36	6	0.0007 ***

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 22: Analysis of Condition 10: Split <2 vs. 6> similar grade yellow-purple

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	0.15	1	0.6966
Scenarios	424.89	6	<2e-16 ***
Advantage: Scenarios	5.77	6	0.4490

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 23: Analysis of Condition 12: <2 vs. 6> dissimilar grade yellow-purple

Factor	LR Chisq	Df	Pr(>Chisq)
Advantage	0.51	1	0.4740
Scenarios	394.66	6	<2e-16 ***
Advantage: Scenarios	1.10	6	0.9813

Signif. codes: '***' p<0.001; '**' p<0.01; '*' p<0.05; '.' p<0.1

Table 24: Study Condition<9>: Split<2 vs. 6>, Grade<similar grade>, Advantage<black-white>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 2 Black - ARP fair 2 Black	-2.143	0.651	-3.293	0.0602942	0%	10%
NDKL fair 2 Black - ARP fair 1 Black	-2.134	0.647	-3.299	0.0592438	0%	10%
NDKL fair 1 White - ARP fair 1 White	1.435	0.400	3.592	0.0227490	60%	32.5%
NDKL fair 1 White - ARP fair 2 White	1.507	0.399	3.780	0.0115424	60%	17.5%
NDKL fair 1 Black - ARP fair 2 Black	2.052	0.430	4.771	0.0001593	47.5%	10%
NDKL fair 2 Black - ARP fair 2 White	-3.074	0.637	-4.828	0.0001202	0%	17.5%
NDKL fair 1 Black - ARP fair 1 Black	2.061	0.424	4.865	0.0000999	47.5%	10%
NDKL fair 2 Black - ARP fair 1 White	-3.146	0.640	-4.917	0.0000769	0%	32.5%
ARP fair 2 Black - NDKL fair 1 White	-2.438	0.433	-5.630	0.0000016	60%	10%
ARP fair 1 Black - NDKL fair 1 White	-2.447	0.426	-5.743	0.0000008	60%	10%

Table 25: Study Condition<10>: Split<2 vs. 6>, Grade<similar grade>, Advantage<yellow-purple>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 1 purple - ARP fair 1 yellow	1.375	0.407	3.379	0.0460966	50%	22.5%
ARP fair 2 purple - NDKL fair 1 yellow	-1.406	0.406	-3.466	0.0347933	50%	22.5%
NDKL fair 2 purple - ARP fair 2 yellow	-1.641	0.464	-3.535	0.0276256	12.5%	15%
NDKL fair 2 purple - ARP fair 1 purple	-1.639	0.463	-3.542	0.0269561	12.5%	17.5%
NDKL fair 2 purple - ARP fair 2 purple	-1.666	0.468	-3.564	0.0250656	12.5%	22.5%
NDKL fair 1 yellow - ARP fair 2 yellow	1.431	0.401	3.569	0.0246263	50%	15%
ARP fair 1 purple - NDKL fair 1 yellow	-1.433	0.400	-3.583	0.0234624	50%	17.5%
NDKL fair 1 purple - ARP fair 2 purple	1.489	0.412	3.613	0.0210939	50%	22.5%
NDKL fair 1 purple - ARP fair 2 yellow	1.515	0.407	3.719	0.0144729	50%	15%
NDKL fair 1 purple - ARP fair 1 purple	1.517	0.406	3.733	0.0137473	50%	17.5%
NDKL fair 2 purple - ARP fair 1 yellow	-1.780	0.465	-3.832	0.0095044	12.5%	22.5%

Table 26: Study Condition<11>: Split<2 vs. 6>,Grade<dissimilar grade>,Advantage<black-white>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 2 White - ARP fair 1 White	-1.560	0.454	-3.434	0.0386285	2.5%	17.5%
NDKL fair 1 White - ARP fair 2 White	1.496	0.402	3.719	0.0144361	40%	15%
NDKL fair 2 Black - ARP fair 2 White	-2.994	0.678	-4.416	0.0008422	5%	15%
NDKL fair 2 Black - ARP fair 1 White	-3.247	0.678	-4.788	0.0001462	5%	17.5%
NDKL fair 1 Black - ARP fair 2 Black	2.268	0.436	5.203	0.0000175	35%	5%
ARP fair 2 Black - NDKL fair 1 White	-2.493	0.438	-5.691	0.0000011	40%	5%
NDKL fair 1 Black - ARP fair 1 Black	2.638	0.456	5.786	0.0000007	35%	5%
ARP fair 1 Black - NDKL fair 1 White	-2.863	0.458	-6.249	0.0000001	40%	5%

Table 27: Study Condition<12>: Split<2 vs. 6>,Grade<dissimilar grade>,Advantage<yellow-purple>

contrast	estimate	SE	z.ratio	p.value	Preference %	
					NDKL	ARP
NDKL fair 1 purple - ARP fair 1 yellow	1.818	0.407	4.468	0.0006647	57.5%	20%
NDKL fair 1 purple - ARP fair 2 purple	1.849	0.405	4.560	0.0004360	57.5%	22.5%
NDKL fair 1 purple - ARP fair 2 yellow	1.876	0.406	4.615	0.0003370	57.5%	12.5%
NDKL fair 1 yellow - ARP fair 1 yellow	1.948	0.400	4.865	0.0001001	70%	20%
NDKL fair 1 purple - ARP fair 1 purple	1.979	0.406	4.872	0.0000967	57.5%	15%
ARP fair 2 purple - NDKL fair 1 yellow	-1.979	0.399	-4.961	0.0000617	70%	22.5%
NDKL fair 1 yellow - ARP fair 2 yellow	2.006	0.400	5.016	0.0000465	70%	12.5%
ARP fair 1 purple - NDKL fair 1 yellow	-2.109	0.400	-5.278	0.0000117	70%	15%

C APPENDIX 3

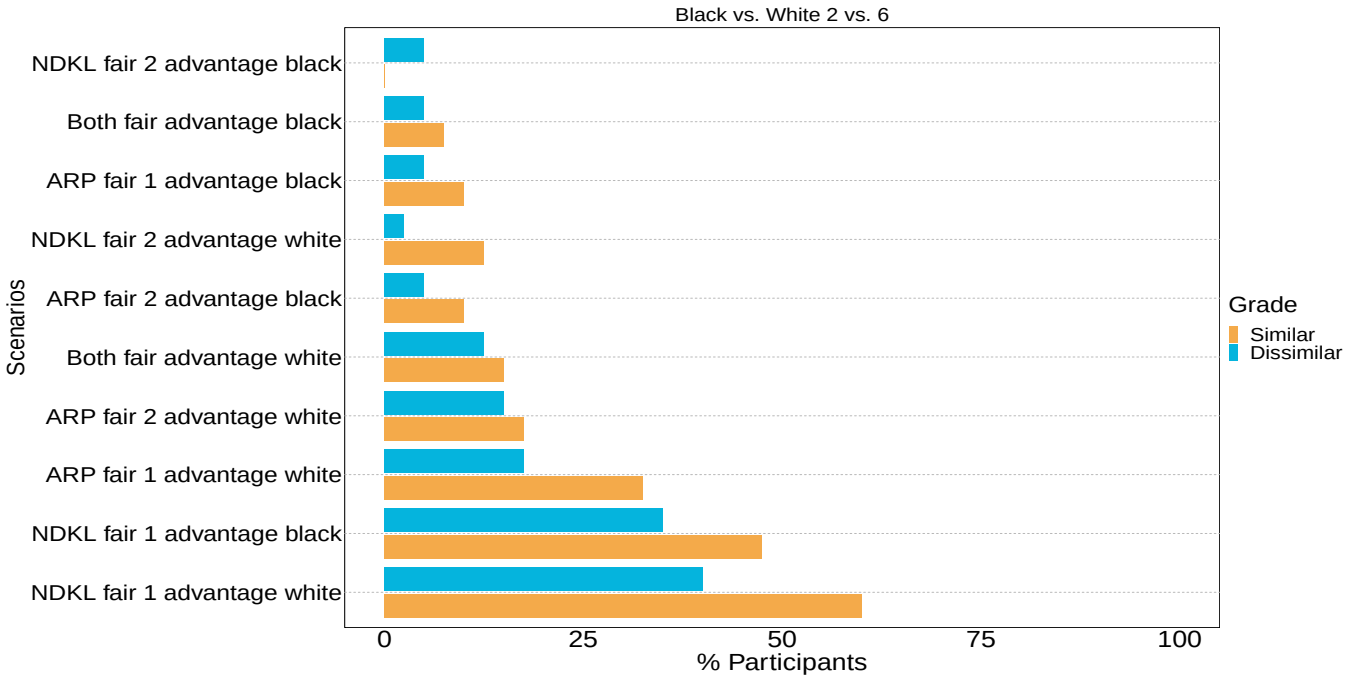

Figure 13: Black vs. White 2 vs. 6 | Grade: Similar and Dissimilar

D APPENDIX 4

Table 28: Comparing Demographic Distribution: Our Prolific vs. 2022 U.S. Census

Demographic Attributes		
Gender		
Attribute	**Prolific (%)**	**Census (%)**
Man	46.67%	49.6%
Woman	49.79%	50.4%
Political Affiliation		
Conservative	11.04%	36%
Liberal	34.79%	37%
Moderate	18.33%	25%
Education		
Associate degree	10.42%	8.8%
Bachelor's degree	34.17%	21.6%
Graduate degree	15.21%	14%
High school	13.96%	26.1%
Some college	25.83%	19.1%
Ethnicity		
American Indian or Alaska Native	2.08%	1%
Asian	7.08%	5.9%
Black or African American	7.71%	12.2%
Native Hawaiian or Other Pacific Islander	0.21%	0.2%
White	78.13%	60.9%
Age		
18-24 years old	19.58%	9.3%
25-34 years old	36.67%	13.6%
35-44 years old	17.50%	13.2%
45-54 years old	13.75%	12.1%
55-64 years old	8.96%	12.7%
More than 64 years old	3.33%	17.3%

E APPENDIX 5

Table 29: This table displays selected quotes collected from our study condition, organized by themes. Every quote is associated with the specific study condition and participant ID, providing a reference to the context and source of the response.

Study Condition	Participants ID	Quote	Theme
1	P1	"Money is not given out based on merit, I don't know what the logic was with this one but I don't like it".	Theme 1
2	P2	"The only fair elements here are that the top scorer did receive the top award. and that in the aggregate each group was awarded an equal amount. However, this committee is obviously racist or sexist or ableist against purple people".	Theme 1
3	P3	"This order is just strange. A 66 is just barely passing, so I wouldn't have put that over the two students that got mid-high 80s".	Theme 1
4	P4	"I think it is unfair because all of the purple group is getting money that should go to yellow group. For example. Robertson is getting money that should go to Uutela. This is causing a huge mess and Simpson and Knight who should be getting significantly more money are getting very little".	Theme 1
5	P5	"Another example of randomly distributing the scholarships. It needs to be based on scores, regardless of the distribution by race".	Theme 1
6	P6	"The two students with the highest grades seem to be receiving random amounts of award money and without any other criteria explaining this decision, it does not make sense".	Theme 1
7	P7	"At this moment I believe regardless of race unless in special circumstances, if your score is below 70... there is no way you should lead a student with a score above 80".	Theme 1
8	P8	"The student in the purple group did not score the highest grade so they should not be awarded the most money".	Theme 1
9	P9	"While the Black students deserve help, this seems to be too much of a penalty to me for the White students".	Theme 1
10	P10	"Award still doesn't scale fairly since A. Johnson is receiving more than the better performing members of the yellow group".	Theme 1
11	P11	"I don't think that K. Rose, who scored better than 6 of his peers, should be getting the third lowest sum of money".	Theme 1
12	P12	"Ranking seems arbitrary, B. Bolen should not be at the bottom with regard to his grade."	Theme 1
1	P13	"This division may not seem completely fair to the students with the higher grades but does somewhat compensate for the differences in socioeconomic disparity".	Theme 2
2	P14	"I think it's a fair distribution based on scoring but might need some adjustment to provide for disadvantage students".	Theme 2
3	P15	"Well, it's fair if the white students have some kind of socioeconomic disadvantage (like coming from a very poor community), but otherwise it's unfair because there seems to be no logical reason to rank white students with lower grades above black students with higher grades".	Theme 2

Table 30: This table displays selected quotes collected from our study condition, organized by themes. Every quote is associated with the specific study condition and participant ID, providing a reference to the context and source of the response.

Study Condition	Participants ID	Quote	Theme
4	P16	"I think I can see what happened - the students in each team are ordinally ranked separately, and those who share an ordinal number are hierarchically organized for award distribution. Those with higher scores eliminate the tie breaker problem, as the higher scorer gets the larger award for their ordinal number. While this makes for a mathematically more equal distribution of prize money, it still leaves something to be desired. What is "fair" in this instance? Knight scored higher than Robinson, but Robinson is getting over 3x what Knight gets. So, if this scholarship is based only on academic achievement, then this method is unfair (assuming their grades are appropriately indicative of their total GPA across their entire academic career). Also, there is no explanation for how socioeconomic situations or minority statuses affect ordinal rankings. It is unwise to assume that the only reason for a lower grade is because class or minority status, so this ranking methodology would be unfair to me. On the other hand, it is indeed a symmetrical sorting system, so it's not illogical....it just needs more transparency. "	Theme 2
5	P17	"This ranking is closer to fair, however you could add a disadvantage weight to bring the minority a bit more equal to the lower majority ranks".	Theme 2
6	P18	"This ranking is based solely upon grading and does not account for other criteria which should be included for a proper assessment and fair distribution of awards. Gender, socio-economic status, race, and even attendance should be assessed".	Theme 2
7	P19	"If the black students have disadvantages like socioeconomic struggles in addition to discrimination of their race, then they should get more scholarship money than the white students. I'd need more info".	Theme 2
8	P20	"Not too fair, James and Reed being boosted a little is arguably understandable given their potentially lower quality of education, but the boost is far too big for reed, it's not fair to any of the team yellows to be ranked below them except maybe duke".	Theme 2
9	P21	"Students are ranked based on their score, except for highest scoring black student which possibly takes into account for socioeconomic differences".	Theme 2
10	P22	"It seems correct to me; unless there are good reason for a different placing of awards, which I do not know about".	Theme 2
11	P23	"More information is needed both the spread seems fair when considering the discrimination Black children experience within the school system and society".	Theme 2
12	P24	"Seems fair but without knowing the background of the students, some may have disadvantages educationally".	Theme 2

F APPENDIX 6

Figure 14: Experiment overview of 4 vs. 4 split.

G APPENDIX 7

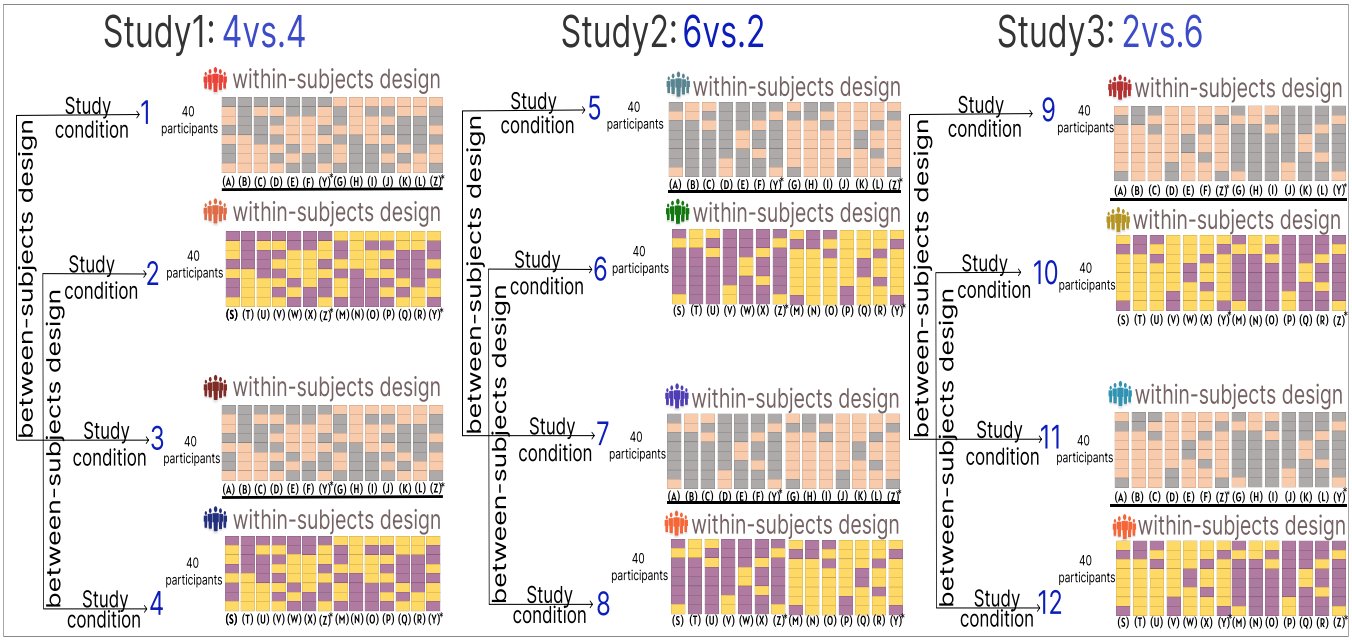

Figure 15: This figure illustrates the mixed-method trial design incorporating both within-subject and between-subject approaches across twelve study conditions. Each condition varies by type of split, grade similarity, and advantage, detailed as follows:

Study Condition <number>: Split <Type of split >, Grade <Type of grade >, Advantage <Type of Advantage >

- Study Condition 1: 4 vs. 4, similar grade, black-white [14 random scenarios]
- Study Condition 2: 4 vs. 4, similar grade, yellow-purple [14 random scenarios]

Link to the survey on Study Conditions 1 and 2.

- Study Condition 3: 4 vs. 4, not similar grade, black-white [14 random scenarios]
- Study Condition 4: 4 vs. 4, not similar grade, yellow-purple [14 random scenarios]

Link to the survey on Study Conditions 3 and 4.

- Study Condition 5: 6 vs. 2, similar grade, black-white [14 random scenarios]
- Study Condition 6: 6 vs. 2, similar grade, yellow-purple [14 random scenarios]

Link to the survey on Study Conditions 5 and 6.

- Study Condition 7: 6 vs. 2, not similar grade, black-white [14 random scenarios]
- Study Condition 8: 6 vs. 2, not similar grade, yellow-purple [14 random scenarios]

Link to the survey on Study Conditions 7 and 8.

- Study Condition 9: 2 vs. 6, similar grade, black-white [14 random scenarios]
- Study Condition 10: 2 vs. 6, similar grade, yellow-purple [14 random scenarios]

Link to the survey on Study Conditions 9 and 10.

- Study Condition 11: 2 vs. 6, not similar grade, black-white [14 random scenarios]
- Study Condition 12: 2 vs. 6, not similar grade, yellow-purple [14 random scenarios]

Link to the survey on Study Conditions 11 and 12.