Exploring "Just Noticeable" Group Fairness in Rankings

Mallak Alkhathlan, Hilson Shrestha, Lane Harrison, Elke Rundensteiner

Worcester Polytechnic Institute, USA
{malkhathlan, hshrestha, ltharrison, rundenst}@wpi.edu

Abstract

The plethora of fairness metrics developed for ranking-based decision-making raises the question: which metrics align best with people's perceptions of fairness, and why? Most prior studies examining people's perceptions of fairness metrics tend to use ordinal rating scales (e.g., Likert scales). However, such scales can be ambiguous in their interpretation across participants, and can be influenced by interface features used to capture responses. We address this gap by exploring the use of two-alternative forced choice methodologies— used extensively outside the fairness community for comparing visual stimuli— to quantitatively compare participant perceptions across fairness metrics and ranking characteristics. We report a crowdsourced experiment with 224 participants across four conditions: two alternative rank fairness metrics, ARP and NDKL, and two ranking characteristics, lists of 20 and 100 candidates, resulting in over 170,000 individual judgments. Quantitative results show systematic differences in how people interpert these metrics, and surprising exceptions where fairness metrics disagree with people's perceptions. Qualitative analyses of participant comments reveals an interplay between cognitive and visual strategies that affects people's perceptions of fairness. From these results, we discuss future work in aligning fairness metrics with people's perceptions, and highlight the need and benefits of expanding methodologies for fairness studies.

1 Introduction

Background on Fairness in Ranking-Based Decisions. AI-driven algorithms ranking candidates in education, healthcare, recruitment, and financial lending risk introducing biases against groups based on race, gender, or disability (Lee 2018; Harris et al. 2023; Loefflad and Grossklags 2024; Zhang 2022; Gadiraju et al. 2023). Given this trend, much attention recently has focused on developing fairness metrics for algorithmic decision-making, and evaluating people's perceptions of fairness in decision-making contexts, emphasizing its role in trust and system adoption (Yurrita et al. 2023; van Berkel, Sarsenbayeva, and Goncalves 2023; Schoeffer, Machowski, and Kuehl 2021; Lee and Rich 2021; Schoeffer, Kuehl, and Machowski 2022).

These works show the alignment of people's perceptions of fairness with fairness metrics in the literature and highlight the importance of transparency in presenting data to users. Across application areas such as legal and criminal risk (Grgic-Hlaca et al. 2018), healthcare (Lee and Rich 2021), and hiring (Lavanchy et al. 2023), ongoing challenges and opportunities in designing fair systems remain. Gaining a deeper understanding of people's perceptions of fairness in rankings across a wide range of tasks will allow organizations to design systems that will deliver equitable and widely accepted outcomes (Lee and Rich 2021).

Gap in the Literature. Unfortunately, the predominant approach for measuring fairness perception relies on qualitative measures and ordered scales, typically Likert scales (Binns 2018; Shulner-Tal, Kuflik, and Kliger 2022; Schoeffer, Kuehl, and Valera 2021; Corbett-Davies et al. 2017; Nyarko, Goel, and Sommers 2021; Marcinkowski et al. 2020; Harrison et al. 2020). These techniques have known biases and limitations, including cultural differences in participant response patterns (Lee et al. 2002; Johnson, Shavitt, and Holbrook 2011), variability in interpretation among respondents (Westland 2022), and tendencies such as acquiescence bias (He et al. 2014; Pimentel and Pimentel 2019; Douven 2018). Further, many of these works (Lee and Rich 2021; Lavanchy et al. 2023) focus on fairness for classification tasks, while ranking tasks have received comparatively less attention.

Our Approach. In this paper we explore an alternative methodology for assessing people's perceptions of fairness namely, the two-alternative forced choice (2AFC) method (Heinrich, Kaur, and O'Donoghue 2015; Derrick, Hansmann, and Theys 2019; Guo et al. 2019), and apply it to study fair-ranking metrics. 2AFC is often used to quantify and model just-noticeable differences (JNDs) for stimuli, which is the minimum difference between two stimuli that an observer can reliably distinguish (Yu and Grauman 2015; Du et al. 2024; Ma et al. 2024; Kay and Heer 2016; Chung et al. 2016; Lu et al. 2022; Harrison et al. 2014). 2AFC and its associated modeling methodologies have been widely adopted to evaluate people's perceptions with the design and evaluation of visual, tactile, and interactive systems (Stevens 1957; Gescheider 1997; Macknik and Martinez-Conde 2009; Xue et al. 2023). In our case, we propose to adopt and adapt it for the purpose of measuring perceptual thresholds in fairness metrics, which we here call "Just-Noticeable Fairness". A key strength of the 2AFC methodol-

ogy is that it effectively mitigates issues related to scale bias (e.g., some participant never selecting 5 or even 4 on a 5-point Likert scale; while others spread their ratings across all 5 scores), as participants only choose between two stimuli. A metric with a consistently lower JND suggests that participants can more easily detect changes in that metric. 2AFC modeling also facilitates examining the impact of contextual variables, such as the size of candidate lists, impacting people's perceptions.

Research Study. We report on our crowdsourced study with 224 participants designed to assess their perceptions of two competing fairness metrics, and to examine the relative impact of candidate list sizes. We compare two alternative metrics for measuring fairness across rankings: ARP (Attribute Rank Parity) (Cachel, Rundensteiner, and Harrison 2022; Shrestha et al. 2022, 2023) and NDKL (Normalized Discounted Cumulative KL-divergence) (Geyik, Ambler, and Kenthapadi 2019; Ghosh, Dutt, and Wilson 2021; Kumar et al. 2023; Olulana et al. 2024), both of which have been applied in multiple recent studies.

We examine **four conditions:** (A) Different Ranking Sizes and Same Metric - ARP_{100} vs. ARP_{20}; (B) Different Ranking Sizes and Same Metric - $NDKL_{100}$ vs. $NDKL_{20}$; (C) Same Ranking Size, and Different Metrics - ARP_{100} vs. $NDKL_{100}$; and (D) Same Ranking Size and Different Metrics - ARP_{100} vs. $NDKL_{100}$. These four conditions explore how fairness metrics and how list sizes influence participants' perceptions of fairness – with their perceptions measured effectively by adopting the 2AFC methodology.

While prior work has examined people's perceptions of visual and visualization stimuli using Just Noticeable Differences (JND) methodologies (Kay and Heer 2016; Harrison et al. 2014), our study is the first to explore this methodology in a *fair ranking context.* Our core research questions about fair-ranking metrics are:

- **RQ1: How does the number of ranked candidates influence people's subjective perceptions of fairness?**
- **RQ2: How do ranking fairness metrics align with people's subjective perceptions of fairness?**

For RQ1, our findings—perhaps surprisingly—indicate that *longer* lists (i.e., rankings with more candidates) led to more consistent and precise fairness judgments from participants. Results clearly show that both ARP_{100} and $NDKL_{100}$ outperformed their 20-candidate counterparts in terms of precision and alignment with fairness perceptions.

For RQ2, the results reveal distinct patterns between ARP and NDKL. ARP more closely aligns with participants' fairness perceptions, consistently yielding lower variance and smaller JNDs for participants to detect fairness differences.

In addition, our qualitative study revealed two main strategies that participants used to judge fairness: cognitive strategies (e.g., intuitive judgments like avoiding clusters of same-group candidates or spotting imbalance, and analytical methods like comparing ratios or distributions) and visual strategies (e.g., focusing on even color distribution, symmetry, and balanced patterns across rankings).

Our findings suggest that ***visual displays*** play a critical role in aiding user perception of fairness. Since our study relies solely on visual representations, it offers insight into when such cues are effective and when they may fall short. This is particularly important in contexts where numeric fairness scores may be inaccessible, not intuitive, or misaligned with users' sense of fairness.

2 Background

2.1 Perception of Algorithmic Fairness

Empirical research on algorithmic fairness has predominantly focused on *classification tasks.* Studies by Srivastava et al. (Srivastava, Heidari, and Krause 2019) and Saxena et al. (Saxena et al. 2019) indicate that non-expert perceptions often align with fairness metrics like demographic parity or calibrated fairness. For instance, Srivastava et al. (Srivastava, Heidari, and Krause 2019) found that demographic parity— where outcomes are equally distributed across groups— aligns closely with laypeople's fairness intuition for classification tasks, while Saxena et al. (Saxena et al. 2019) reveal a preference for calibrated fairness, where resources like money are divided according to applicants' quality, with perceptions influenced by factors such as race. Debjani et al. (Saha et al. 2020) developed a metric to assess non-expert comprehension of fairness definitions, highlighting gaps between algorithmic fairness and public perception and the need for transparent explanations.

While classification studies focus on fairness in discrete outcomes, ranking fairness (Yang and Stoyanovich 2017; Zehlike et al. 2017; Zehlike and Castillo 2020; Singh and Joachims 2018) adds complexity by addressing the relative ordering of items and the distribution of protected groups across a list (Lee et al. 2019; Ötting and Maier 2018; Wang, Harper, and Zhu 2020; Lavanchy et al. 2023). Unlike classification, which uses metrics like accuracy, ranking fairness must account for how item positions affect perceptions and outcomes, such as scholarship amounts, raises, or other benefits (Kuhlman and Rundensteiner 2020; Wu et al. 2022; Sühr, Hilgard, and Lakkaraju 2021; Diaz et al. 2020).

Research on AI explanations has shown that transparency and the way information is framed significantly influence fairness perceptions. For example, Schoeffer et al. (Schoeffer, De-Arteaga, and Kühl 2024) found that clear explanations foster trust, while Zhang et al. (Yang et al. 2024) developed tools to mitigate bias. However, Binns et al. (Binns et al. 2018) noted that certain explanation styles, like case-based explanations, can negatively affect fairness views. Other studies have highlighted how user characteristics, such as education, gender, or perceived system reliability, also shape how fairness is judged (Pierson 2017; Srivastava, Heidari, and Krause 2019; Saxena et al. 2019; Saha et al. 2020; Lünich and Keller 2024).

Similar to our work, Alkhathlan et al. (Alkhathlan et al. 2024) used Likert scales to study perception of fair rankings, however, they examined the influence of score distributions of relevant attributes (e.g., similar vs. dissimilar grades in a scholarship scenario) on fairness perception. Further, their study was restricted to small lists composed of only 8 candidates, while most real-world problems must work much larger sets of candidates that need to be fairly ranked.

Building on this, our research now investigates how list size (short vs. long) affects fairness perceptions in ranking systems. Unlike prior fairness perception studies, we explicitly introduce 2AFC methodologies to calculate Just Noticeable Differences that allow us to capture precise differences in how people perceive fairness (Yu and Grauman 2015; Du et al. 2024; Ma et al. 2024; Kay and Heer 2016; Chung et al. 2016; Lu et al. 2022; Harrison et al. 2014).

2.2 Assessing JND Using 2AFC Methods

The concept of Just Noticeable Difference (JND) refers to the smallest detectable difference between two stimuli perceived by the human sensory system (Yu and Grauman 2015; Du et al. 2024; Ma et al. 2024; Kay and Heer 2016; Chung et al. 2016; Lu et al. 2022) Lu et al. (Lu et al. 2022) quantified perceptible differences in visual attributes like bar height and bubble size to enhance data visualizations. Chung et al. (Chung et al. 2016) explored how visual channels, such as size and texture, affect order perception. Ma et al. (Ma et al. 2024) extended the concept of JND to wearable devices like AirPush, which provides multi-dimensional force feedback for immersive VR. Du et al. (Du et al. 2024) explored tactile feedback through surface roughness. Studies by Harrison and Kay et al. (Harrison et al. 2014; Kay and Heer 2016) model correlation perception in visualizations using Weber's Law, while Yu et al. (Yu and Grauman 2015) explore the perceptibility of differences in visual attributes sportiness in images. These studies show how JND-focused methods can inform the design of interactive systems.

The development of visualization systems relies deeply on empirical studies of human visual perception, which also benefit efforts in human-computer interaction (HCI). Rensink et al. explored the perception of correlation in scatterplots using psychophysical methods to assess visual perception in data visualization (Rensink and Baldridge 2010). This work modeled how visual attributes scatterplot shape might affect data interpretation. Harrison et al. extended this method to multiple chart types, demonstrating how model-based approaches can be used to compare competing methods (Harrison et al. 2014). Kay et al. introduced advanced statistical methods to the modeling approach, incorporating individual differences and hierarchical models, and adding uncertainty measurement and modeling (Kay and Heer 2016).

JNDs in different contexts are estimated by studies using a Two-Alternative Forced Choice (2AFC) method (Heinrich, Kaur, and O'Donoghue 2015; Derrick, Hansmann, and Theys 2019; Guo et al. 2019). 2AFC is used to assess perceptible differences in stimuli by asking participants to choose between two options, enabling the quantification of Just Noticeable Differences (JND) (Szafir 2017; Elliott et al. 2021). While prior work has used pairwise comparisons for preference elicitation—such as in participatory voting systems (Goel et al. 2019). Our work draws on psychophysical evaluation frameworks as proposed by Elliott et al. (Elliott et al. 2021), which emphasize sensitivity, threshold, and performance-based measures in visualization assessment. Compared to Likert-style ratings, 2AFC allows for finer-grained sensitivity to perceptual thresholds in fairness judgments. Szafir applied 2AFC to measure color differences in visualizations, aligning design with human perceptual thresholds (Szafir 2017). Elliott et al. focuses on stimulus adjustments and performance measures to improve visualization evaluation (Elliott et al. 2021).

3 Methodology

Overall Goal: Finding the perceptual threshold. Our study uses a controlled perception-based evaluation approach adopted from the visualization research community to assess how fairness is perceived by ("looks" to) people when they are shown ranked list displays. By isolating visual perception from broader social context, the goal is to establish *perceptual thresholds*—the minimum fairness differences participants can reliably detect between two rankings with different levels of fairness. For example, if one ranking has a fairness score F_1 (e.g., 0.2, interpreted as mostly fair) and another has F_2 (e.g., 0.9, interpreted as quite unfair), the perceptual threshold represents the smallest difference at which participants can detect a fairness difference between the two displayed lists. In this example, the difference is $|F_2 - F_1| = |0.9 - 0.2| = 0.7$, if participants are able to perceive it.

A small perceptual threshold (small just noticeable difference, or JND) indicates that participants can detect even minor fairness changes, while a large threshold would mean that even sizable fairness differences may go unnoticed. This experimental methodology therefore allows us to examine whether participants can visually distinguish between different fairness objectives—independent of context or specific candidate demographics (e.g. race or sex).

Fairness Metrics. To compute the fairness F_i of a list, we utilize two popular group fairness metrics in rankings that align with the concept of statistical parity (Dwork et al. 2012), which seeks proportional representation of protected groups across ranked positions. These metrics, Normalized Discounted KL-divergence **(NDKL)** (Geyik, Ambler, and Kenthapadi 2019) and Attribute Rank Parity **(ARP)** (Cachel, Rundensteiner, and Harrison 2022), were selected for their distinct approaches to measuring group fairness in the sense of prioritizing fairness at different parts of the ranking as elaborated upon below.

NDKL assesses the representation of groups at every prefix of the ranking, weighting higher-up prefixes more. It deems a ranking fair if each prefix, i.e., a top-k set, 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. The metric is defined as:

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

where $d_{KL}(D_{\tau_i} || D_X)$ measures the KL-divergence between the group proportions at the first i positions in τ and the overall group distribution in X. The normalization factor Z ensures comparability across rankings and is defined as

$Z = \sum_{i=1}^{|\tau|} \frac{1}{\log_2(i+1)}$. Higher NDKL values indicate greater disparities in representation at top positions.

ARP is a pairwise fairness metric that evaluates rankings by decomposing them into pairwise comparisons, where a mixed pair compares candidates from different groups. ARP measures fairness by calculating the maximum absolute difference between the average mixed pairs won by each group. Fairness is achieved when ARP equals 0, indicating that groups are treated equally across the entire ranking. The ARP metric is defined as:

$$ARP(\tau, G) = \max_{\forall g_j, g_k \in G} |\text{avgpairs}(\tau, g_j) - \text{avgpairs}(\tau, g_k)| \quad (2)$$

where avgpairs(τ, g_i) is calculated as the ratio of mixed pairs won by group g_i to total mixed pairs involving that group. Higher ARP values, up to 1, suggest that groups are disproportionately ranked lower compared to others.

3.1 Study Design

Our study examines the impact of *two core dimensions*—competing fairness metrics (ARP and NDKL) and alternate list sizes (small = 20, large = 100)—yielding four higher-level comparison conditions (A–D), summarized in Figure 3. The rationale for focusing on these two dimensions is discussed in the subsections entitled *Varying Fairness Metrics* and *Varying Ranked List Sizes* below.

Participants were presented with two rankings of different fairness levels (according to some metric F) and asked to choose the ranking they perceived as being "more fair" (See study interface in Figure 1.) We call this a trial. A ranking corresponds to an ordered sequence of candidates, each with a sensitive attribute denoting their membership in a protected group (e.g., the two groups sex = male or sex = female). For simplicity, a binary group system is assumed: half of the candidates belong to Group A and the other half to Group B, visually indicated by yellow and purple respectively in our study interface (See Figure 1). We chose yellow and purple as abstract socially-neutral colors denoting different groups to avoid triggering biases, following prior findings that participants respond differently when explicit categories (e.g., race) are used compared to abstract ones (e.g., colors) (Alkhathlan et al. 2024).

As described below in the Fairness Metrics subsection, we worked with two fairness metrics from the literature: ARP and NDKL. As second dimension of our study, we automatically generated rankings for each metric in *two list sizes:* 20 candidates and 100 candidates. These rankings are denoted as *Metric*$_X$, where *X* represents the list size, either 20 or 100, and Metric the abbreviation for the fairness metric. For example, ARP_{100} and NDKL_{100} correspond to the longer lists, while ARP_{20} and NDKL_{20} refer to the shorter lists.

Participant Assignment. Participants were randomly assigned to one of four *studies:* ARP_{100}, ARP_{20}, $NDKL_{100}$, or $NDKL_{20}$, through Prolific (see the Recruitment and Participants subsection below for participant requirements). Each participant is assigned only one study (See study interface in Figure 1). This between-group design ensured that each participant was exposed to only one combination of fairness metric and list size (e.g., ARP_{100}). To illustrate, participants in the ARP_{100} study only viewed trials comparing ARP_{100} rankings to other ARP_{100} rankings; they were not exposed to rankings by other metrics or list sizes to avoid confusion.

All participants completed the same set of trials, as described in the following procedure and main study subsections. Each trial presents a unique pair of rankings for comparison, meaning no pair of rankings was repeated across trials within the same study. Trials were independent, such that participants' responses to one trial did not influence the presentation or outcomes of other trials. Participants had to complete all trials. The study design ensured balanced representation by including an equal number of trials per condition within each study, and by keeping the total number of trials consistent across the four studies.

The trial involved presenting a pair of rankings, where the fairness score for each ranking (e.g., using ARP or NDKL) varied systematically across predefined *fairness levels:* $(0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)$ (See Figure 2C.i). If an exact fairness level (e.g., 0.3) was not achievable for a given list size and metric combination (e.g., NDKL_{20}), we used the closest possible fairness score instead (e.g., 0.26).

More precisely, each ranking (also called the *first ranking*) at a given base fairness level (e.g., level 0.5) is paired with multiple other rankings. As shown in Figure 2C.ii, it is paired with 24 other rankings, referred to as the *second rankings*. The pairings were constructed with fairness differences ranging from 0.02, 0.04, 0.1, etc. from the base fairness value to detect both small and larger differences. These trials (i.e., tasks where participants select the fairer ranking between two side-by-side rankings, with one correct answer) explore a range of comparisons, capturing the metric's responsiveness to subtle or large differences in fairness. This includes when *first ranking* might be completely fair (e.g., a score of 0.0), moderately fair (e.g., a score of 0.5), or not fair at all (e.g., a score of 1.0), with the *second ranking* reflecting contrasting levels of fairness. Their position (left vs. right) in the pair is randomized.

The two rankings (*first* and *second*) never have the same fairness score, as participants are always required to choose which ranking is "more fair" without an option to indicate no difference. To avoid ordering bias, the order of trials presented to participants was randomized within each study.

We evaluate perceptions of fairness across 194 trials for each of the four studies: ARP_{100}, NDKL_{100}, ARP_{20}, and NDKL_{20}, totaling $4 \times 194 = 776$ trials. With 224 subjects, this yielded $776 \times 224 = 173{,}824$ individual judgments.

Varying Fairness Metrics. Fairness metrics differ in how they evaluate representation across ranked lists. ARP (Cachel, Rundensteiner, and Harrison 2022) measures fairness by comparing overall group proportions. For example, in a ranking [F, M, F, M, F, M] (where F = female, M = male), ARP indicates high fairness (ARP = 0.33) due to balanced counts. In contrast, NDKL (Geyik, Ambler, and Kenthapadi 2019) penalizes disparities more at higher ranks. Under NDKL, a ranking [M, M, M, F, F, F] would be consid-

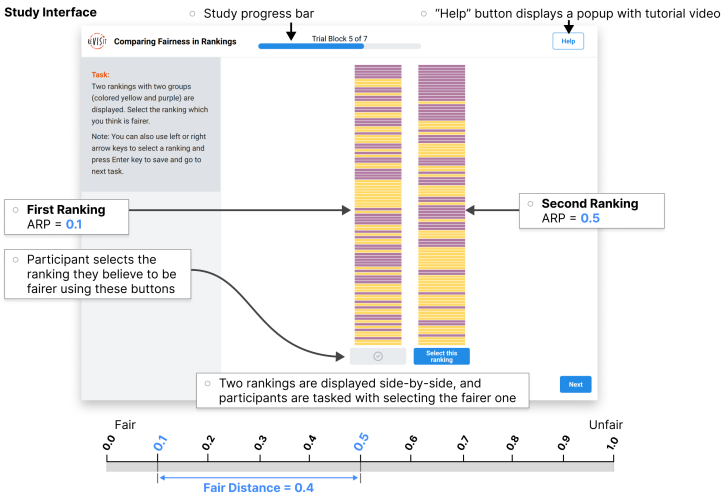

Figure 1: The study interface for one of our four studies (e.g., ARP_{100}) illustrates two rankings side-by-side with different fairness scores. Participants are tasked with selecting the ranking they believe is "more fair." The actual metric scores are not shown to participants. In this illustration, a ranking with ARP = 0.5 is paired against a ranking with ARP = 0.1, yielding a fairness metric distance of 0.4.

ered less fair (NDKL = 1.0) because the underrepresented group appears lower. Our study uses both ARP and NDKL to examine how different fairness metrics highlight distinct fairness concerns in rankings.

Varying Ranked List Sizes. Verma et al. (Verma et al. 2023) explored how group size disparities influence fairness in resource allocation, finding that fairness decreased with significantly disproportionate group sizes (e.g., 3 vs. 120 beneficiaries) due to compassion fade (Slovic et al. 2007) and drop-in-the-bucket effects (Bartels and Burnett 2011). Inspired by this, our study examines whether increasing the number of candidates in a ranking from 20 to 100 affects fairness perception and potentially decision-making, assessing how these processes adapt to larger scenarios.

Aggregation of Studies. We aggregated the data from the four studies ARP_{100}, ARP_{20}, $NDKL_{100}$, and $NDKL_{20}$ into four higher-level comparison conditions (A–D) to address our research questions, as detailed below and shown in Figure 3. This aggregation allowed us to analyze fairness perceptions across metrics and ranking sizes by combining results from participants exposed to metrics ARP and NDKL, and to short (20 candidates) and long (100 candidates) lists, as follows:

(A) Different Ranking Sizes and Same Metric – ARP_{100} vs. ARP_{20}: In Condition A, the ARP metric is applied to both a short list of 20 candidates and a long list of 100 candidates. This condition A examines how list size affects fairness perception using the same metric ARP.

(B) Different Ranking Sizes and Same Metric – $NDKL_{100}$ vs. $NDKL_{20}$: In Condition B, the NDKL metric is applied to both a short list of 20 candidates and a long list of 100 candidates. Condition B explores the influence of list size on fairness perception using a different metric (namely, NDLK) than Condition A.

(C) Same Ranking Size and Different Metrics – ARP_{100} vs. $NDKL_{100}$: Condition C compares the perceived fairness of ARP and NDKL for a short list of 20 candidates, isolating the effect of the metric type on fairness judgments.

(D) Same Ranking Size and Different Metrics - ARP_{100} vs. $NDKL_{100}$: In Condition D, ARP and NDKL are compared using a 100 candidate list to see how the choice of metric affects fairness perception on larger lists.

3.2 Study Procedure

The study was structured as follows.

Tutorial Video. Before beginning the study, all participants watched a required one-minute tutorial video that introduced the concept of fairness in ranked decisions. This video used a concrete real-world example—such as women and men receiving unequal outcomes in scholarship allocation—to illustrate group fairness. By framing the ranking task in familiar domains like hiring or education, the tutorial ensured that participants had a shared understanding of fairness in social decision-making before proceeding with the trials (see Figure 2, Part A). The training video is available in supplemental material.

Practice Session. In this practice, participants completed 12 practice trials of varying difficulty (see Figure 2, Part B) to become familiar with the task format. Each study condition included a matched training session, which used examples aligned with the fairness metric and ranking size used in that condition. For example, participants in the ARP-20 condition saw ARP-based rankings with 20 candidates. Trials were divided into three difficulty levels: easy (e.g., scores near 0 and 1), fair distance $\approx$ 1), medium (e.g., scores of 0.2 and 0.9), and hard (e.g., scores of 0.02 and 0.04). Each session included 6 easy, 3 medium, and 3 hard trials. After

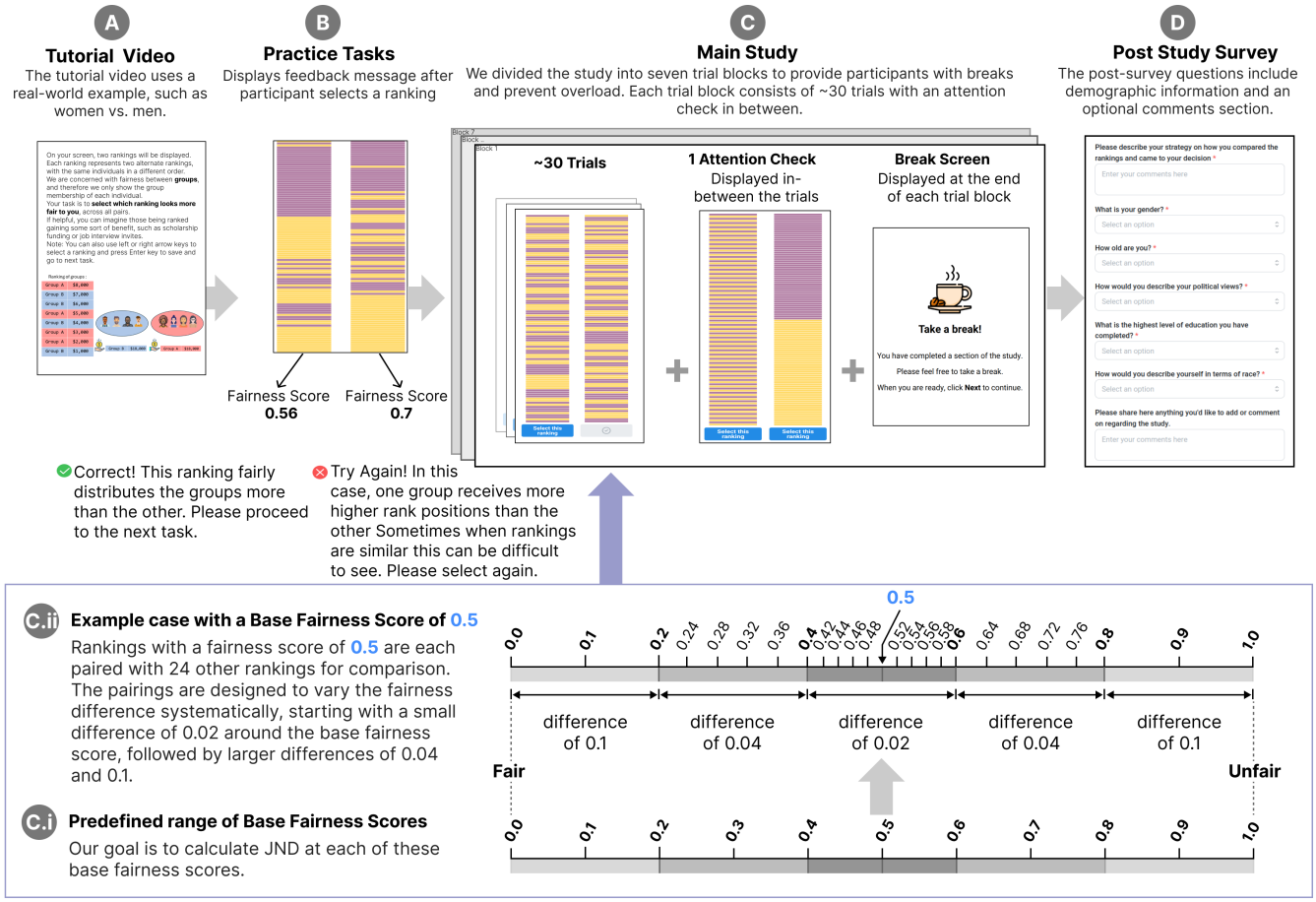

Figure 2: Participants first completed (A) a tutorial video to introduce the concept of fairness in ranked decisions and provide domain context, followed by (B) practice tasks with feedback, (C) main study, and (D) a post-study survey. Supplementary materials, including *Short Video Overview*, *Tutorial Video*, *Experiment Code*, and *Data Analysis Code*, are available at https://osf.io/drs67/.

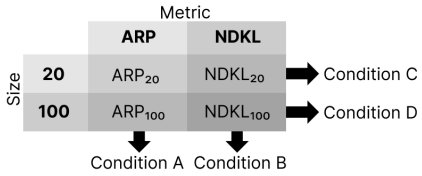

Figure 3: We study four higher-level comparison conditions (A–D) across the combination of ranking sizes (20 candidates vs. 100 candidates) and metrics (ARP vs. NDKL).

each trial in this practice, participants received feedback on whether their answer was correct or incorrect. This feedback was not included in the main study.

Main Study. To manage participant workload and maintain focus, we divided each main study of 194 trials assigned to a participant into seven *blocks*, with each block comprising a group of trials. Blocks 1 through 6 each contained 30 trials, and block 7 contained the remaining 14 trials, for a total of $(6 \times 30) + 14 = 180 + 14 = 194$ trials per participant. Each block included one randomly placed attention check. These checks resembled—but did not exactly match—the *Easy Trials* from the practice session. A short break screen was shown after each block to reduce fatigue.

Post-Study Data Collection. At the end of the experiment, participants provided demographic information (gender, age, political views, education level, race) and responded to open-ended questions (Figure 2D) to share optional comments, feedback, and describe the strategies they used to compare rankings and make decisions. This post-study assessment allowed us to gather both demographic data and qualitative insights into participants' thought processes and fairness judgments.

3.3 Recruitment and Participants

We determined the number of participants for each condition through pilot studies and power analyses. We adopt methods from prior studies that fit models using pilot data (e.g. (Yang et al. 2023)), and use posterior draws from a Bayesian model to simulate larger numbers of participants. Bayesian power

simulations, conducted by refitting the model for each condition, indicated that 60 participants achieved reliable performance, with an Effective Sample Size (ESS) of 458 and an Rhat value of approximately 1.002.

We conducted the study using the Prolific platform (Prolific n.d.) for participant recruitment, employing the reVISit software framework to deploy web-based studies (Ding et al. 2023). The participant pool filters were set to individuals at least 18 years old, in the U.S., with reported high English proficiency, and a minimum approval rating of 100. An IRB-approved informed consent was given. Participant completion and payment was determined through automatic recording of Prolific IDs via URL parameters. Participants were compensated at a rate of $12 per hour, following an estimated median completion time of the study being 22.80±6.25 minutes.

The total number of participants across all conditions was 255. From these, 31 participants either failed attention checks, or completed the study too rushed to be trustworthy (in under 5 minutes), leaving 224 participants for the analysis. Reported gender distribution includes 74 men, 145 women, and 5 non-binary. Most participants identified as White (165) and Black or African American (44), with smaller groups identifying as Asian (9), American Indian or Alaska Native (2), and Native Hawaiian or Other Pacific Islander (1); three chose not to disclose their race. Reported age includes 20 participants 18-24, 69 aged 25-34, 46 aged 35-44, 56 aged 45-54, 21 aged 55-64, and 12 over 64. Political views vary, with 45 conservatives, 64 liberals, 51 moderates, 28 somewhat conservative, 35 somewhat liberal, and 1 who preferred not to respond.

3.4 Modeling

Dependent Variable: Correctness. This binary dependent variable indicates whether a participant correctly answered a question, with 1 for correct and 0 for incorrect responses. It serves as the primary measure to assess participants under various fairness levels. Given the binary nature of 2AFC correctness, we adopt logistic regression models, as they estimate the probability of a participant responding correctly and are suitable for comparisons both within metrics *Fair Distance* and across metrics.

Independent Variable: Fair Distance. As the primary predictor, *Fair Distance* (Figure 1) measures the actual distance between two stimuli in terms of the fairness metrics (Figure 2B). This estimates how *Fair Distance* impacts the probability of correct responses, and allows adjustments to be made for interactions with other model variables.

Model Fitting Using Analysis with BRMS. Our Bayesian logistic regression model, specified using the *brm()* function in R, predicts the probability of selecting the fairer ranking based on *Fair Distance* (Figure 1), the numerical difference in fairness scores between two pair rankings. The model accounts for variability across participants and *Base Fairness Levels* (Figure 1) by incorporating random slopes for *participant* and *Base Fairness Level*, allowing the effect of *Fair Distance* on the outcome (*Correctness*) to vary across participants and fairness levels.

4 Results and Analysis

To answer the two key research questions, we analyze the variation in JND values for conditions A, B, C, and D in Figures 4, 5, 6, and 7, respectively, comparing people's perceived fairness of different ranking metrics across fairness levels from complete fairness 0.0 to total unfairness 1.0.

4.1 RQ1: How Does the Number of Ranked Candidates Influence People's Perceptions of Fairness?

Condition (A): Different Ranking Sizes and Same Metric – ARP_{100} vs. ARP_{20}.

All levels: In condition (A)(Figure 4),we compared ARP for two list sizes: 100 candidates versus 20 candidates. At the base fairness level of 0.0 (completely fair), ARP_{100} has a lower JND threshold, indicating a more stable perception of fairness compared to ARP_{20}. Overall, ARP_{20} tends to show larger JNDs and variance than ARP_{100}.

At fairness level 0.1, ARP_{100} has a small JND threshold of 0.07, continuing to outperform ARP_{20}. At fairness level 0.2, ARP_{20} shows similar performance to ARP_{100}.

At mid-range fairness levels, ARP_{100} maintains smaller JND thresholds, ranging between 0.12 and 0.15. As fairness levels approach 1.0 (complete unfairness), ARP_{100} achieves its lowest JND values of 0.15 at 0.8, 0.07 at 0.9, and 0.05 at 1.0, demonstrating more stable fairness perceptions under higher unfairness.

Example 1: At fairness level 0.0 (completely fair), ARP_{100} achieved a JND threshold of 0.10, with 98% of participants correctly identifying the fairness ranking between 0.26 and 0.0. Conversely, ARP_{20} reported a higher JND of 0.39, with 89% accurately recognizing the ranking pair between 0.06 and 0.0. These results highlight how ranking size affects fairness perception. (Figure 4, Example Trial 1).

Condition (B): Different Ranking Sizes and Same Metric –$NDKL_{100}$ vs. $NDKL_{20}$

All levels: (Figure 5), we compared NDKL for two list sizes: 100 candidates versus 20 candidates. At the base fairness level of 0.0, considered completely fair, $NDKL_{20}$ achieves a JND of 0.10, indicating robust performance and a stable perception of fairness. Similarly, at the fairness level of 1.0, seen as completely unfair, both $NDKL_{100}$ and $NDKL_{20}$ maintain a JND of 0.10, demonstrating consistent perception regardless of list size. $NDKL_{100}$ consistently shows lower JND thresholds, ranging from 0.10 to 0.21 across all levels, reflecting its strong stability in fairness perception. The smallest JND of 0.08 at fairness level 0.9 highlights $NDKL_{100}$'s exceptional stability at higher levels of perceived unfairness.

At fairness 0.1, $NDKL_{100}$ reaches the highest JND value of 0.21, indicating a decrease in stability. $NDKL_{20}$ shows a JND of 0.24 at the same level, suggesting less stability. At fairness levels 0.3, 0.4, and 0.6, $NDKL_{100}$ maintains a JND range of 0.12 to 0.16, underscoring stability, while $NDKL_{20}$ shows higher variability with JNDs up to 0.56 at level 0.4.

Example 2: At fairness level 0.4, $NDKL_{100}$ achieved a JND of 0.15, with 74% of participants correctly identifying the ranking pair 0.4 and 0.36. In contrast, $NDKL_{20}$ reported

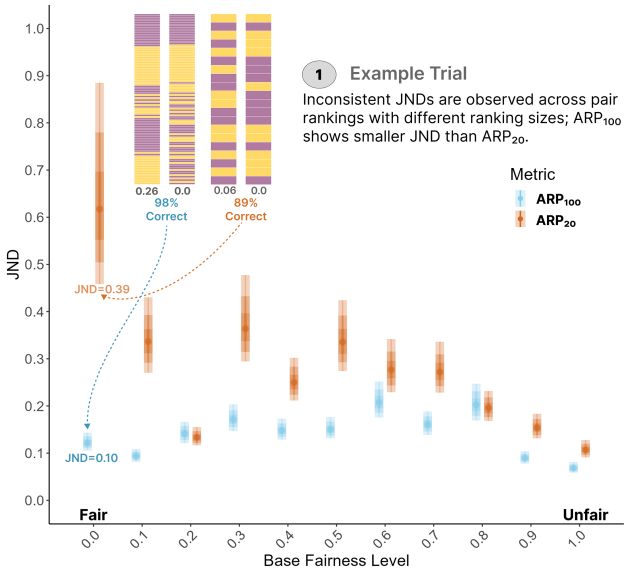

Figure 4: **Condition (A):** Comparison of ARP_{100} and ARP_{20}. Longer lists with ARP tend to yield more precise fairness judgments, with few exceptions.

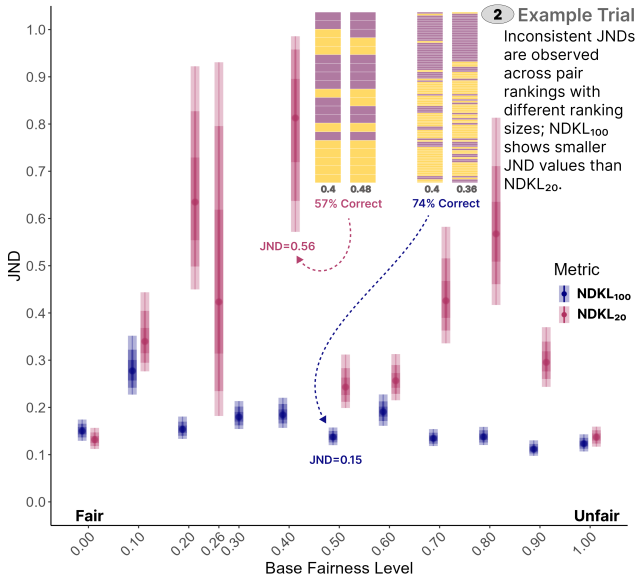

Figure 5: **Condition (B):** $NDKL_{100}$ vs. $NDKL_{20}$. Smaller lists showed greater disagreement and higher JNDs, while larger lists improved consistency.

a higher JND of 0.56, with 57% correctly recognizing the ranking pair 0.4 and 0.48. These results highlight how list size impacts fairness perception. (Figure 5, Example trial 2).

Key takeaways. Larger lists improve fairness perception. In Conditions *A* and *B*, both ARP_{100} and $NDKL_{100}$ show stronger and more stable perception than their 20-item versions. ARP_{100} also yields more noticeable fairness differences, supporting the use of longer rankings.

4.2 RQ2: How Do Ranking Fairness Metrics Align With People's Subjective Perceptions of Fairness?

Condition (C): Same Ranking Size and Different Metrics – ARP_{20} vs. $NDKL_{20}$

All levels: In condition (C) (Figure 6), we compared ARP_{20} and $NDKL_{20}$ for a short list of 20 candidates. At the baseline fairness level of 0.0 (fair), $NDKL_{20}$ achieves a low JND threshold of 0.10, indicating a stable perception of fairness under ideal conditions. However, as fairness levels increase from 0.1 to 0.3, ARP_{20} shows lower JND thresholds, suggesting greater sensitivity to slight unfairness.

One interesting divergence appears at fairness level 0.4, where $NDKL_{20}$ experiences a marked increase in JND threshold 0.56, whereas ARP_{20} remains notably lower, demonstrating its stronger performance in detecting moderate deviations from fairness. At mid-range fairness levels (0.5–0.6), $NDKL_{20}$ recovers some stability (JND thresholds of 0.17 and 0.20, respectively), yet ARP_{20} continues to show superior sensitivity overall.

At fairness level 0.7, ARP_{20} again outperforms $NDKL_{20}$, maintaining a lower JND threshold of 0.21 compared to $NDKL_{20}$'s 0.28. Moving toward the higher end of unfairness (0.8–1.0), ARP_{20} sustains its advantage with JND thresholds of 0.11 at 0.8 and 0.9, and an even lower threshold of 0.08 at 1.0, indicating that people perceive fairness more reliably when using ARP_{20} in highly unfair contexts.

Example 3: At a fairness level of 0.4 (medium fairness), $NDKL_{20}$ had a JND threshold of 0.56, with 21% of participants identifying the 0.8 vs. 0.4 ranking pair. In contrast, ARP_{20} achieved a lower threshold of 0.19, with 96% recognizing the 0.4 vs. 0.14 pair, showing how each metric shapes fairness perception. (See Figure 6, Example 3.)

Condition (D): Same Ranking Size and Different Metrics – ARP_{100} vs. $NDKL_{100}$

All levels: In condition (D) (Figure 7), we compared ARP_{100} and $NDKL_{100}$. At the baseline fairness level (0.0), ARP_{100} demonstrates a JND threshold of 0.10, suggesting effective fairness detection under ideal conditions. As unfairness increases, ARP_{100} continues to perform robustly with consistently lower JND thresholds, particularly noticeable at fairness levels of 0.1, 0.2, 0.3, and 0.4 with thresholds of 0.07, 0.11, 0.13, and 0.12, respectively, indicating its heightened sensitivity to subtle fairness deviations.

Conversely, $NDKL_{100}$ presents more variable JND thresholds across the fairness range, performing comparably in mid-range and slightly unfair scenarios. It registers JND thresholds of 0.11 at fairness levels 0.5, 0.7, and 0.8, and a slightly higher threshold of 0.15 at fairness level 0.6, suggesting challenges in consistently gauging fairness.

The performance gap is most pronounced at fairness level 0.4, where ARP_{100} achieves a lower JND threshold of 0.12 compared to $NDKL_{100}$'s 0.15, confirming ARP_{100} as more reliable in detecting moderate unfairness. At fairness extremes, both metrics converge, with ARP_{100} performing best at complete unfairness (1.0) with a JND threshold of 0.05, highlighting its effectiveness in clear unfair scenarios.

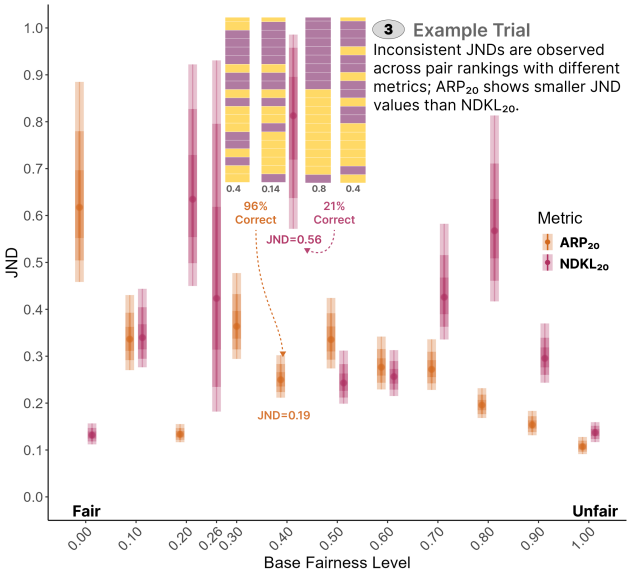

Figure 6: **Condition (C):** Comparison of ARP_{20} and $NDKL_{20}$. Smaller lists yield higher JNDs and variance. ARP_{20} shows lower values overall, indicating more reliable perceived differences.

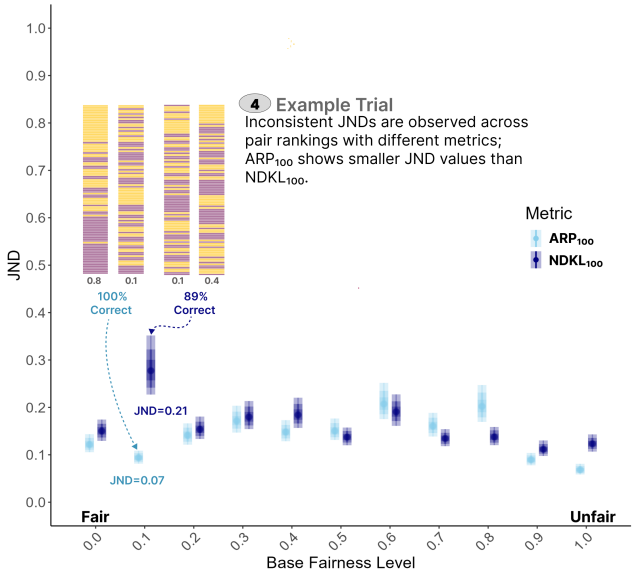

Figure 7: **Condition (D):** Comparison of ARP_{100} and $NDKL_{100}$. Larger lists yield lower JNDs and variance.

Example 4: At fairness level 0.1 (mostly fair), ARP_{100} achieved a JND threshold of 0.07, with 100% of participants correctly identifying the fairness ranking between 0.8 and 0.1. Conversely, $NDKL_{100}$ reported a higher JND threshold of 0.21, with 89% recognizing the ranking pair between 0.1 and 0.4. These results illustrate how metric choice influences fairness perception. See Figure 7, Example 4.

Key takeaway. In both Conditions (C) and (D), ARP aligns more closely with participants' fairness perceptions than NDKL, showing consistently lower JNDs across fairness levels. This suggests ARP is more effective and reliable for supporting perceived fairness in both short and long lists.

Observations at Extreme Fairness Levels: As shown in Table 1, not surprisingly, the smallest JND values often appear at the extreme fairness levels 0.0 (completely fair) and 1.0 (completely unfair), along with their adjacent points 0.1 and 0.9. Among these extremes, ARP_{100} typically achieves the lowest JNDs (e.g., 0.05 at fairness 1.0), indicating strong alignment with participants' perceptions of fairness in both fair and unfair scenarios. Not surprisingly, participants align more closely with fairness metrics at fairness extremes.

Level	ARP_{100}	ARP_{20}	$NDKL_{100}$	$NDKL_{20}$
0.0	**0.10**	0.39	0.12	**0.10**
0.1	**0.07**	0.23	0.21	0.24
0.9	**0.07**	0.11	0.08	0.22
1.0	**0.05**	0.08	0.10	0.10

Table 1: JND for fairness metrics at extreme fairness levels.

4.3 Qualitative Analysis

Participants also provided open-response explanations, sharing their thoughts and reasoning during each condition. We employed a systematic open-coding methodology to identify main and subthemes in this qualitative data. Analyis followed a reflexive thematic approach, where three members of the research team independently coded the responses and then collaboratively refined the themes (Byrne 2022; Morgan 2022). This iterative process helped capture the nuanced perspectives and recurring patterns in the data. To assess inter-rater reliability, we used Cohen's kappa statistic, with agreement scores ranging from 0.8 to 1. This high level of consensus among coders strengthens the reliability of our insights. Overall, we analyzed 224 feedback entries, which reflected participants' strategies for selecting fair rankings.

Cognitive Strategies: This theme includes the subthemes *"Intuitive Judgments"* and *"Analytical Methods"*. Participants relied on decision-making processes or a mix of both quantitative and qualitative reasoning to determine fairness.

Under the subtheme *"Intuitive Judgments"*, participant $NDKL_{P40}$ drew on their experience stating, *"It reminded me of doing annual 'stack ranking' of my employees at Microsoft... I would always focus on 'tops and tails' because ... So here, I focused on top/bottom and looked for overt clustering there that would imply bias or a 'bad manager' did the ranking."* This shows how their judgment, informed by past experience, focussed on extremes in rankings. Similarly, participant $NDKL_{P42}$ stated, *"I tried my best to determine which side looked most balanced. I went with my first instinct ... I will say some looked the same or equal, but the colors were just moved around."*. These examples illustrate how participants often relied on gut feelings to assess fairness, aligning with the cognitive strategy of intuitive judgment when evaluating apparent imbalances in rankings.

In contrast, the subtheme *"Analytical Methods"* highlights more methodical approaches. Participant $NDKL_{P63}$ explained, *"To determine which bar had a 50/50 split, I started by closely examining the data for each bar, focusing on the number of items or outcomes in each category... I calculated the ratios for each bar by dividing the smaller count by the larger count, aiming for a result close to 0.5."* This participant's decision-making process was based on calculating the balance between two categories to ensure fairness, providing a clear quantitative strategy. ARP_{P27} took a more qualitative stance, remarking, *"To me, the bars represented socio-economic opportunity with higher levels more valuable than each subsequent level. I considered that top levels had more value (e.g., greater value of access to business or home loans)."*. These examples highlight a structured approach with contextual reasoning to perceive fairness.

Visual Perceptions: This theme includes the subthemes *"Color Distribution"* and *"Balanced Patterns"*. Participants focused on visual cues such as the evenness of the color distribution and the overall balance of rankings.

Under the subtheme *"Color Distribution"*, ARP_{P2} shared, *"I attempted to determine which selection had the widest distribution of the colors that were spread over more area"* showing a preference for visually diverse rankings. $NDKL_{P10}$ similarly noted, *"I tried to pick the side that had the most even distribution of colors"* while ARP_{41} added, *"I mostly looked at how spread out the colors were, and whether there were long sections of one color at a time."*. These examples highlight fairness being perceived through the even spread of colors across rankings.

For *"Balanced Pattern"*, participant ARP_{P18} stated, *"Which pattern was most balanced and evenly distributed."* Likewise, $NDKL_{P19}$ shared, *"I just went off of looks, if one side was heavily favored towards the top I tried to pick the one that brought the other side in as close to the top as possible."* Meanwhile, $NDKL_{P5}$ took a more imaginative approach, saying, *"I tried to imagine that if I were to place a fulcrum between the 10th and 11th row and that each colored square were given a weight of 1 unit, then I should assume the correct selection is the distribution that is more evenly balanced on a theoretical scale."*. These instances demonstrate how fairness was perceived through balance and symmetry in rankings, with participants focusing on even weight distribution and avoiding imbalances.

5 Discussion

Fairness and Ranking Data Characteristics. Our findings show that the number of ranked candidates influences people's perception of fairness. Larger lists (e.g., 100 candidates) lead to more consistent JND values and a more reliable understanding of fairness. The abundance of data in longer rankings may create a more visually discernible difference between fairness levels, especially when each candidate is color-coded by group. This effect may arise from candidates of the same group appearing in denser clusters, making patterns of disparity more noticeable.

Towards Building More Transparent Decision-Support Systems. Our results have potential implications for designers, policymakers, and researchers developing decision-making tools in that while ***visual displays*** are powerful, their effectiveness can vary depending on context and user needs. As such, designers of fairness-aware decision-support systems should consider when to complement visuals with numeric indicators (e.g., fairness scores) or explanatory guidance. These insights are especially relevant in domains like hiring, education, or credit decisions, where fairness must not only be achieved but also clearly conveyed. By identifying perceptual thresholds and boundaries, our work may help inform the design of more transparent AI-assisted systems.

6 Limitations and Future Work

This study's participant pool was limited to U.S.-based respondents and was demographically imbalanced—most participants identified as White and female—which may have influenced fairness perceptions. While this provided some cultural consistency, it constrains generalizability. Future work should recruit more diverse and internationally representative samples to better capture how fairness perceptions vary across demographic and cultural contexts.

We used abstract visual stimuli with color-coded group indicators to isolate perceptual thresholds in 2AFC tasks. While this design enabled focus on general perceptual differences, it may not generalize to real-world settings with explicit group identities. The ranked lists were vertical and visually encoded; though effective, this reflects only one format. Future work should explore alternative layouts, orientations, and encodings.

We focused on two fairness metrics—ARP and NDKL. Our findings suggest limitations in distinguishing between fairness levels when differences are subtle (e.g., in $NDKL_{20}$). Future work should evaluate additional metrics and extend beyond binary group settings to include three or more protected groups. Subgroup analyses by race, gender, or socioeconomic status may further clarify how different communities interpret fairness.

Finally, while our findings highlight which fairness differences are perceptible, their practical implications remain exploratory. Future research should examine how these thresholds can inform fairness-aware system design in applied contexts.

7 Conclusion

This study examined how popular group fairness metrics align with laypeople's fairness perceptions through a crowdsourced experiment with 224 participants. Using 2AFC to estimate just noticeable differences in fairness levels from visually-displayed ranked decisions, we found that longer lists improved participants' ability to discern fairness differences, and that ARP generally aligned more closely with fairness perceptions. Results suggest that cognitive and visual cues—such as intuitive judgments and visual patterns—shape fairness perception. These findings offer insights into when a decision maker is likely to 'notice' fairness—or its absence—in a visual ranking, highlighting when intuitive cues are sufficient and when formal fairness metrics may be needed for fairness to be perceptible.

Ethical Statement

Considerations Statement

In this study, we represented sensitive groups and membership using abstract constructs (e.g., colors such as Yellow and Purple) to avoid associating groups with specific demographic or identity markers. This design choice aimed to reduce participants' reliance on stereotypes or biases, allowing them to focus solely on evaluating fairness in rankings.

We intentionally excluded other individual characteristics beyond group membership in the actual data collection portion of the study to minimize confounding factors. However, during the practice session, participants were presented with generalized contexts (e.g., scholarship funding, interview invitations) to provide context and relevance of this work, while avoiding bias tied to a particular application domain.

All participants received a detailed consent form outlining study objectives, their rights, and the voluntary nature of participation, with the option to withdraw at any time. A "Prefer not to respond" option was included in demographic data collection to respect participant privacy. Personally identifiable information was collected solely for payment purposes and kept separate from study data. Prolific IDs were anonymized, and demographic data were aggregated and removed prior to analysis. All researchers completed CITI Program training in human subjects research and data security.

Researcher Positionality Statement

We approach this research as scholars trained in human-computer interaction, computational fairness, and as computer scientists, drawing on perspectives shaped by academic and Western understandings of fairness, equity, and procedural justice. Our choice to abstract group identities reflects a commitment to reducing stereotype activation in experimental settings, yet we recognize this decision is itself situated within specific research and cultural paradigms.

We acknowledge that our framing of fairness—as something evaluable through participant perceptions of ranked outcomes—reflects our positionality and is not neutral. We welcome critical engagement with our design choices and interpretations.

Adverse Impact Statement

While this study uses abstract representations to reduce bias, we acknowledge that abstraction may risk oversimplifying fairness by detaching it from structural inequities tied to race, gender, or other identities. This could inadvertently suggest that fairness can be fully captured by mathematical metrics, independent of social and historical contexts.

Additionally, the absence of explicit demographic identifiers may limit the study's relevance to individuals with lived experiences of discrimination.

Finally, while our displays did not explicitly encode fairness cues, this should not be interpreted as opposing transparency in AI systems. We affirm the importance of disclosing fairness metrics and methodologies to promote accountability and trust.

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).

References

Alkhathlan, M.; Cachel, K.; Shrestha, H.; Harrison, L.; and Rundensteiner, E. 2024. Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1940–1970.

Bartels, D. M.; and Burnett, R. C. 2011. A group construal account of drop-in-the-bucket thinking in policy preference and moral judgment. *Journal of Experimental Social Psychology*, 47(1): 50–57.

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, 149–159. PMLR.

Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; and Shadbolt, N. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, 1–14.

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

Cachel, K.; Rundensteiner, E.; and Harrison, L. 2022. MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 1124–1137. Kuala Lumpur, Malaysia: IEEE.

Chung, D. H. S.; Archambault, D.; Borgo, R.; Edwards, D. J.; Laramee, R. S.; and Chen, M. 2016. How Ordered Is It? On the Perceptual Orderability of Visual Channels. *Comput. Graph. Forum*, 35(3): 131–140.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, 797–806. New York, NY, USA: Association for Computing Machinery. ISBN 9781450348874.

Derrick, D.; Hansmann, D.; and Theys, C. 2019. Tri-modal speech: Audio-visual-tactile integration in speech perception. *The Journal of the Acoustical Society of America*, 146(5): 3495–3504.

Diaz, F.; Mitra, B.; Ekstrand, M. D.; Biega, A. J.; and Carterette, B. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 275–284.

Ding, Y.; Wilburn, J.; Shrestha, H.; Ndlovu, A.; Gadhave, K.; Nobre, C.; Lex, A.; and Harrison, L. 2023. reVISit: Supporting Scalable Evaluation of Interactive Visualizations. In *2023 IEEE Visualization and Visual Analytics (VIS)*, 31–35. IEEE.

Douven, I. 2018. A Bayesian perspective on Likert scales and central tendency. *Psychonomic bulletin & review*, 25: 1203–1211.

Du, X.; Satriadi, K. A.; Drogemuller, A.; Matthews, B. J.; Smith, R.; Walsh, J. A.; and Cunningham, A. 2024. That's Rough! Encoding Data into Roughness for Physicalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: Association for Computing Machinery. ISBN 9781450311151.

Elliott, M.; Nothelfer, C.; Xiong, C.; and Szafir, D. A. 2021. A Design Space of Vision Science Methods for Visualization Research. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 1117–1127.

Gadiraju, V.; Kane, S.; Dev, S.; Taylor, A.; Wang, D.; Denton, E.; and Brewer, R. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 205–216. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Gescheider, G. A. 1997. *Psychophysics: The Fundamentals*. New York: Psychology Press, 3rd edition edition. ISBN 9780203774458.

Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 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*, KDD '19, 2221–2231. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

Ghosh, A.; Dutt, R.; and Wilson, C. 2021. When Fair Ranking Meets Uncertain Inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, 1033–1043. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380379.

Goel, A.; Krishnaswamy, A. K.; Sakshuwong, S.; and Aitamurto, T. 2019. Knapsack Voting for Participatory Budgeting. *ACM Trans. Econ. Comput.*, 7(2).

Grgic-Hlaca, N.; Redmiles, E. M.; Gummadi, K. P.; and Weller, A. 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*, WWW '18, 903–912. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450356398.

Guo, L.; Weems, J. T.; Walker, W. I.; Levichev, A.; and Jaramillo, S. 2019. Choice-selective neurons in the auditory cortex and in its striatal target encode reward expectation. *Journal of Neuroscience*, 39(19): 3687–3697.

Harris, C.; Johnson, A. G.; Palmer, S.; Yang, D.; and Bruckman, A. 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): 1–31.

Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; and Ur, B. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 392–402. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Harrison, L.; Yang, F.; Franconeri, S.; and Chang, R. 2014. Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12): 1943–1952.

He, J.; Espinosa, A. D.; van de Vijver, F.; and Poortinga, Y. 2014. Acquiescent and socially desirable response styles in cross-cultural value surveys. In *Toward sustainable development through nurturing diversity*, 98–111. International Association for Cross-Cultural Psychology.

Heinrich, J.; Kaur, S.; and O'Donoghue, S. 2015. Evaluating the Effectiveness of Color to Convey Alignment Quality in Macromolecular Structures. In *2015 Big Data Visual Analytics (BDVA)*, 1–8.

Johnson, T. P.; Shavitt, S.; and Holbrook, A. L. 2011. Survey Response Styles Across Cultures. In *Cross-Cultural Research Methods in Psychology*, 130–175. New York, NY, US: Cambridge University Press. ISBN 978-0-521-75842-0.

Kay, M.; and Heer, J. 2016. Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1): 469–478.

Kuhlman, C.; and Rundensteiner, E. 2020. Rank aggregation algorithms for fair consensus. *Proc. VLDB Endow.*, 13(12): 2706–2719.

Kumar, D.; Grosz, T.; Rekabsaz, N.; Greif, E.; and Schedl, M. 2023. Fairness of recommender systems in the recruitment domain: an analysis from technical and legal perspectives. *Frontiers in big Data*, 6: 1245198.

Lavanchy, M.; Reichert, P.; Narayanan, J.; and Savani, K. 2023. Applicants' fairness perceptions of algorithm-driven hiring procedures. *Journal of Business Ethics*, 1–26.

Lee, J. W.; Jones, P. S.; Mineyama, Y.; and Zhang, X. E. 2002. Cultural differences in responses to a Likert scale. *Research in nursing & health*, 25(4): 295–306.

Lee, M. K. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1): 2053951718756684.

Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 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): 1–26.

Lee, M. K.; and Rich, K. 2021. Who is included in human perceptions of AI?: Trust and perceived fairness around

healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.

Loefflad, C.; and Grossklags, J. 2024. How the Types of Consequences in Social Scoring Systems Shape People's Perceptions and Behavioral Reactions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1515–1530. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Lu, M.; Lanir, J.; Wang, C.; Yao, Y.; Zhang, W.; Deussen, O.; and Huang, H. 2022. Modeling Just Noticeable Differences in Charts. *IEEE Transactions on Visualization and Computer Graphics*, 28(1): 718–726.

Lünich, M.; and Keller, B. 2024. Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions. *Conference on Fairness, Accountability and Transparency*.

Ma, Y.; Xie, T.; Zhang, P.; Kim, H.; and Je, S. 2024. AirPush: A Pneumatic Wearable Haptic Device Providing Multi-Dimensional Force Feedback on a Fingertip. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Macknik, S. L.; and Martinez-Conde, S. 2009. Real Magic: Future Studies of Magic Should Be Grounded in Neuroscience. *Nature Reviews Neuroscience*, 10(3): 241–241.

Marcinkowski, F.; Kieslich, K.; Starke, C.; and Lünich, M. 2020. Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 122–130. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

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

Nyarko, J.; Goel, S.; and Sommers, R. 2021. Breaking Taboos in Fair Machine Learning: An Experimental Study. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450385534.

Olulana, O.; Cachel, K.; Murai, F.; and Rundensteiner, E. 2024. Hidden or Inferred: Fair Learning-To-Rank with Unknown Demographics. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1088–1099.

Ötting, S. K.; and Maier, G. W. 2018. The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior*, 89: 27–39.

Pierson, E. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.

Pimentel, J. L.; and Pimentel, J. 2019. Some biases in Likert scaling usage and its correction. *International Journal of Science: Basic and Applied Research (IJSBAR)*, 45(1): 183–191.

Prolific. n.d. Prolific – Trusted by Researchers. Accessed: [date].

Rensink, R. A.; and Baldridge, G. 2010. The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, 29(3): 1203–1210.

Saha, D.; Schumann, C.; Mcelfresh, D.; Dickerson, J.; Mazurek, M.; and Tschantz, M. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 8377–8387. PMLR, Online: PMLR.

Saxena, N. A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. C.; and Liu, Y. 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*, AIES '19, 99–106. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Schoeffer, J.; De-Arteaga, M.; and Kühl, N. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Schoeffer, J.; Kuehl, N.; and Machowski, Y. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. *Conference on Fairness, Accountability and Transparency*.

Schoeffer, J.; Kuehl, N.; and Valera, I. 2021. A Ranking Approach to Fair Classification. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '21, 115–125. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384537.

Schoeffer, J.; Machowski, Y.; and Kuehl, N. 2021. A study on fairness and trust perceptions in automated decision making. *arXiv preprint arXiv:2103.04757*.

Shrestha, H.; Cachel, K.; Alkhathlan, M.; Rundensteiner, E.; and Harrison, L. 2022. FairFuse: Interactive Visual Support for Fair Consensus Ranking. In *2022 IEEE Visualization and Visual Analytics (VIS)*, 65–69. IEEE.

Shrestha, H.; Cachel, K.; Alkhathlan, M.; Rundensteiner, E.; and Harrison, L. 2023. Help or Hinder? Evaluating the Impact of Fairness Metrics and Algorithms in Visualizations for Consensus Ranking. *Conference on Fairness, Accountability and Transparency*.

Shulner-Tal, A.; Kuflik, T.; and Kliger, D. 2022. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology*, 24(1): 2.

Singh, A.; and Joachims, T. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD Inter-*

national Conference on Knowledge Discovery & Data Mining, KDD '18, 2219–2228. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Slovic, P.; Finucane, M. L.; Peters, E.; and MacGregor, D. G. 2007. The affect heuristic. *European journal of operational research*, 177(3): 1333–1352.

Srivastava, M.; Heidari, H.; and Krause, A. 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*, KDD '19, 2459–2468. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

Stevens, S. S. 1957. On the Psychophysical Law. *Psychological Review*, 64(3): 153–181.

Sühr, T.; Hilgard, S.; and Lakkaraju, H. 2021. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 989–999. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Szafir, D. A. 2017. Modeling color difference for visualization design. *IEEE transactions on visualization and computer graphics*, 24(1): 392–401.

van Berkel, N.; Sarsenbayeva, Z.; and Goncalves, J. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies*, 170: 102954.

Verma, A.; Morais, L.; Dragicevic, P.; and Chevalier, F. 2023. Designing Resource Allocation Tools to Promote Fair Allocation: Do Visualization and Information Framing Matter? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Wang, R.; Harper, F. M.; and Zhu, H. 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*, CHI '20, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

Westland, J. C. 2022. Information loss and bias in likert survey responses. *PloS one*, 17(7): e0271949.

Wu, H.; Mitra, B.; Ma, C.; Diaz, F.; and Liu, X. 2022. Joint multisided exposure fairness for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 703–714.

Xue, Y.; Jin, J.; Sun, W.; and Lin, W. 2023. HVS-inspired adversarial image generation with high perceptual quality. *J. Cloud Comput.*, 12(1).

Yang, K.; and Stoyanovich, J. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450352826.

Yang, M.; Arai, H.; Yamashita, N.; and Baba, Y. 2024. Fair Machine Guidance to Enhance Fair Decision Making in Biased People. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Yang, X.; Yang, B.; Tang, C.; Mo, X.; and Bin, H. 2023. Visual Attention Quality Research for Social Media Applications: A Case Study on Photo Sharing Applications. *International Journal of Human-Computer Interaction*.

Yu, A.; and Grauman, K. 2015. Just Noticeable Differences in Visual Attributes. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, 2416–2424. USA: IEEE Computer Society. ISBN 9781467383912.

Yurrita, M.; Draws, T.; Balayn, A.; Murray-Rust, D.; Tintarev, N.; and Bozzon, A. 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*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 1569–1578. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349185.

Zehlike, M.; and Castillo, C. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of the web conference 2020*, 2849–2855.

Zhang, M. 2022. Affirmative Algorithms: Relational Equality as Algorithmic Fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 495–507. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.