Help or Hinder? Evaluating Fairness Metrics and Algorithms in Visualization Systems for Consensus Ranking

by

Hilson Shrestha

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

April 2023

APPROVED:

Professor Lane Harrison, Thesis Advisor

Professor Jacob Whitehill, Thesis Reader

Professor Craig A. Shue, Department Head

Abstract

For applications where multiple stakeholders provide recommendations, a fair consensus ranking must not only ensure that the preferences of rankers are well represented, but must also mitigate disadvantages among socio-demographic groups in the final result. However, there is little empirical guidance on the value or challenges of visualizing and integrating fairness metrics and algorithms into human-in-the-loop systems to aid decision-makers. In this work, we design and develop a system called FairFuse, that includes visual encodings of fairness metrics and fair-rank generation algorithms to generate fair consensus rankings. We design a study to analyze the effectiveness of integrating such fairness metrics-based visualization and algorithms. We explore performance through a task-based crowdsourced experiment comparing FairFuse with a similar system for constructing a fair consensus rankings without the inclusion of fairness metrics visualization and algorithms, called ConsensusFuse. We analyze metrics of fairness, agreement of rankers' decisions, and user interactions in constructing the fair consensus ranking across these two systems. In our study with 200 participants, results suggest that providing fairness-oriented support features nudges users to align their decision with the fairness metrics while minimizing the tedious process of manually having to amend the consensus ranking. We discuss the implications of these results for the design of next-generation fairness oriented-systems, along with emerging directions for future research.

Acknowledgements

I would like to thank my advisors, Prof. Lane Harrison and Prof. Elke Rundensteiner, for their guidance, encouragement, and mentorship. I would also like to thank my colleagues Kathleen Cachel and Mallak Alkhathlan, for their hard work, and collaboration. Additionally, I would like to thank my friends and family for their unending love, support, and inspiration. Thank you all from the bottom of my heart.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ubiquitous task of combining preferences by multiple stakeholders into a consensus is challenging for decision-makers that steer this process. Decision-makers often grapple with diverging preferences provided by different stakeholders, and must reach a single decision that all stakeholders accept and agree with. A frequent approach to such decision-making is to employ rankings, where each stakeholder provides their ranking over the candidates. Candidates might include lists of people, organizations, or other entities. Decision-makers combine these base rankings from individual stakeholders into a single consensus ranking as part of the process.

However, when ranking candidates, stakeholders may provide biased or unfair rankings [1]. Bias can be implicit (unintended), for example, when favoring candidates from a particular university who happen to be overwhelmingly white. Bias can also be explicit, for example, weighing women candidates lower due to a perceived lack of ability for the target role.

One way to mitigate such unfair outcomes is by promoting measures from the algorithmic fairness community, such as *group fairness* or *statistical parity* [2]. Statistical parity, for example, is a requirement that all groups receive an equal proportion of the positive outcome; in our case, favorable positions in the consensus

ranking. Without intervention in the ranking process, there is substantial risk of perpetuating unfair practices, and thus harming marginalized groups.

1.1 State of the Art

Constructing a consensus ranking is challenging [3, 4] and ensuring that this consensus ranking is fair is even more difficult [5, 6]. Numerous visualization tools have explored the design space of rankings [7, 8] and rank-based decision making [9, 10, 11]. But existing approaches have not dealt with the complications of incorporating fairness into visual encodings, nor with interactive workflows related to consensus rank generation. A number of interactive visualization systems have been developed to mitigate socio-demographic biases [12, 13, 14, 15, 16, 17, 18, 19, 20]. These systems are either limited to classifications or single ranking. Similarly, while research in fair algorithms has developed rank-focused auditing metrics and fair rank aggregation methods [5, 6], they have been confined to (non-visual) algorithmic solutions requiring substantial technical expertise to use. Studies have also been conducted to evaluate the capabilities of toolkits that incorporate fairness into their process [21, 22, 13, 23]. However, these studies are either limited to ML practitioners or limited number of users.

1.2 Proposed Solution

To address this gap, we contribute the design and development of **FairFuse**, an interactive visualization system for generating, analyzing, and auditing fair consensus rankings. We conduct a goal and activity analysis (*e.g.* [10, 11, 24] to define fairness-oriented tasks in ranking contexts (Table 3.1, 3.2). We propose a parallel-coordinates style visualization design for rankings with a focus on the group membership of

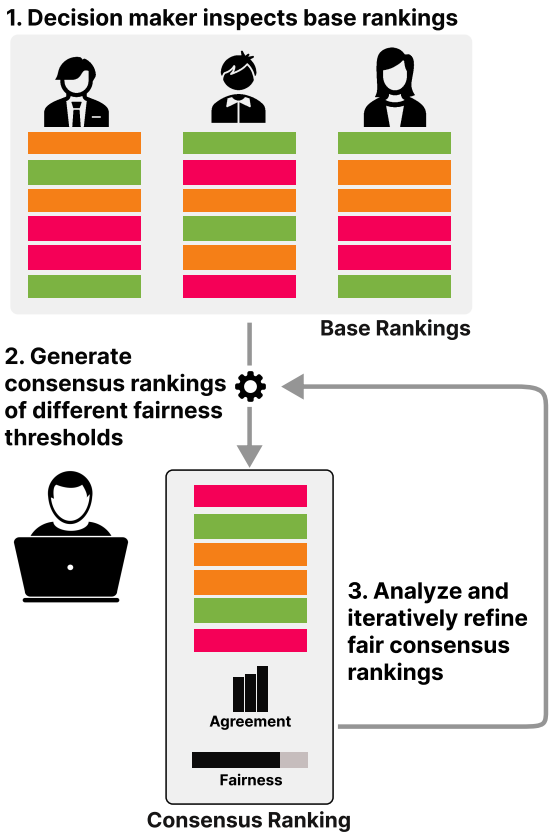


Figure 1.1: Illustrating a fairness-oriented ranking workflow enabled by FairFuse

candidate attributes. We develop novel visual encodings for group-based fairness metrics. FairFuse enables an iterative ranking- and fairness-oriented workflow, allowing decision-makers to visually inspect and edit consensus rankings as part of their decision-making process. Our use cases demonstrate how a decision-maker can use FairFuse in fairness-oriented ranking scenarios.

This research, along with the recent efforts such as work from Mitchell *et al.* and Crisan *et al.* on model cards [25, 26], and Van Berkel *et al.* on examining the value of visualization over text for communicating fairness concepts [27] highlight the inherent promise of interactive visualization for advancing goals in the fairness community. However, efforts combining visualization approaches and fairness metrics and algorithms raise both challenges as well as unique opportunities in this space.

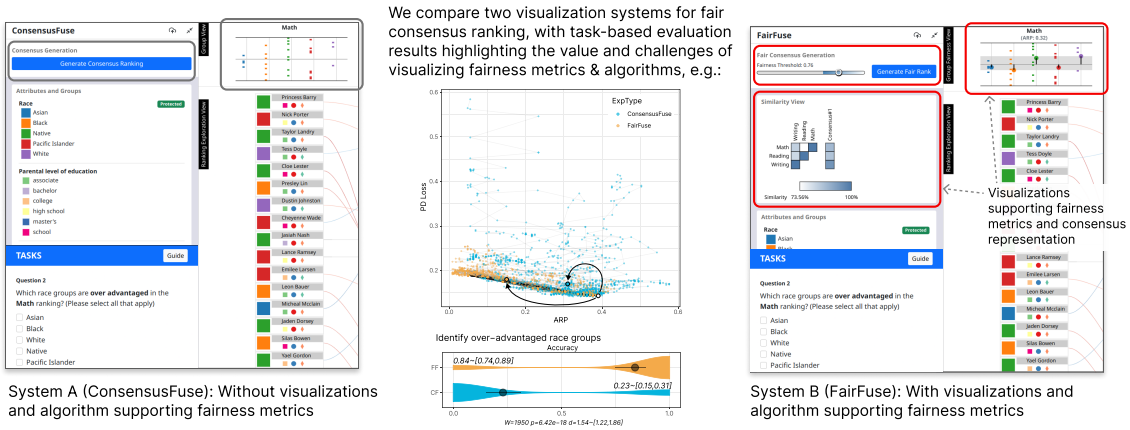


Figure 1.2: Overview of the study. We compare two systems: A: ConsensusFuse, a visualization that enables fairness comparison only by interactive visual displays of underlying items. B: FairFuse, a similar visualization which visualizes additional fairness metrics and provides a fair-rank generation algorithm.

Can visualizations aid some fairness-related tasks, but hinder others? Should fairness metrics be visualized by tightly integrating them with the underlying data items, or separately through popular visualization techniques such as coordinated multiple views? Might some visualizations even mislead or otherwise reduce the agency of users in achieving fairness in decision-making contexts?

In this work, we additionally explore these broad questions through a particular instance of a controlled task-based visualization study (Figure 1.2). We adapt FairFuse into a new system, ConsensusFuse, by removing visualized fairness metrics and algorithms (Figure 3.1). We distill the goals and activities into a set of evaluation tasks, with measurable outcomes (Table 4.1). With the two systems and above identified tasks in place, we conduct a between-subjects task-based evaluation with $n = 200$ participants.

Results generally validate that visualizing fairness metrics leads to notably increased accuracy in key fairness-related tasks, 82% compared to 47% overall. However, deeper analysis of measures, exploration behavior, and participant explanations reveal nuance, challenges and risks in visualizing fairness metrics. In the discussion

section, we review findings, such as the fact that the presence of algorithm-driven fairness schemes tended to "shift" participants' exploration and ultimate decisions in a ranking task. We also develop a set of takeaways highlighting where visualized fairness generally tends to help, but also where it may hinder users in decision-making contexts.

1.3 Contributions

Taken together, this work makes the following contributions:

- Design of an interactive system that visualizes fairness metrics and algorithms to build fair consensus rankings.
- Use cases that demonstrate how a decision-maker can use FairFuse in fairness-oriented ranking scenarios
- A task-based evaluation comparing the system that visualizes fairness metrics and algorithm results against a control with equivalent functionality, sans metrics/algorithms.
- Results that generally validate the value of visualizing fairness metrics and algorithms for rank-focused contexts.
- Additional analyses that highlight particular challenges in visualization design for fairness, including risks and tensions in fairness interface design that may require substantial future effort to resolve.

Chapter 2

Related Works

2.1 Tools and Evaluation Studies on Consensus Building

Visualization systems have been designed to aid decision-makers in inspecting multiple stakeholders' preferences to reach a consensus decision [28, 29, 30, 31, 32, 33, 9, 34, 35, 36, 37, 38]. A subset of these tools consider the setting, like ours, in which stakeholder preferences are encoded as rankings [10, 29, 9]. Liu *et al.* [9] evaluated a between-subjects experiment to assess the effectiveness of their proposed tool, ConsensUs, designed for multiple stakeholders to rate and select candidates. They found that visualizations helped surface stakeholder disagreement that otherwise would have gone undetected. Hindalong *et al.* [11] perform an evaluation study of six tools (both visualization-focused systems and commercial systems that implicitly allow for stakeholder preference inspection), including the systems of [10, 9, 29]. The corresponding evaluation studies are focused on how well these tools help achieve consensus outcomes – yet none consider the employment of consensus generation algorithms [39, 40, 41, 42]. In contrast, we study consensus building when decision makers are supported by fair consensus rank generation algorithms and when fairness metrics are presented visually throughout the process.

2.2 Tools for Ranking-based Tasks and Corresponding Evaluation Studies

Interactive systems and evaluation studies of visualization paradigms have been developed specifically for ranking data. Gratzl *et al.* [7] propose a visualization system, LineUp, to compare ranked items along multiple attributes. Their qualitative evaluation study found that visualizations helped people perform challenging ranking-based tasks faster. Wall *et al.* [8] presented Podium, a visual analytics tool for helping users define a ranking function combining multiple criteria according to their interactions with a subset of the ranked data. Behrisch *et al.* [43] presented a visual system to compare similarities and differences of pairs of rankings using small multiple views of glyphs. However, while the above works target rank-oriented workflows, they neither consider the problem of visually comparing a consensus ranking vis-a-vis the stakeholder's respective base rankings nor how fairness metrics should be incorporated visually throughout the consensus ranking process.

2.3 Visualizing and Presenting Fairness in Information Systems

Much of the work in algorithmic fairness in recent years has focused on proposing various conceptualizations of fairness, along with algorithmic techniques for ensuring these definitions are met in decision-making processes. Comparatively less work has proposed fairness-oriented visualization systems or studied the merits of visual representations of fairness and bias in decision-making.

Fairness Visualization Tools and Toolkits

The design of interactive or visual systems has predominately focused on highlighting and providing recourse for socio-demographic bias in classification tasks [12, 15, 16, 17, 13]. The focus on classification-based machine learning models mirrors the attention of the larger algorithmic fairness community, namely, where “Fair-ML" gained prominence in the context of binary classification. Many tools have been developed to detect algorithmic biases and to evaluate and compare different machine learning models concerning fairness [44, 14, 45]. Crisan *et al.* and Mitchell *et al.* [26, 25] proposed visual model cards for documenting models for better transparency. Recent visualization research has focused on addressing group bias discovery and the interpretation of intersectional bias [46, 47]. In the context of rankings, Yang *et al.* [18] provided “nutritional facts" for the fairness of rankings, Ahn *et al.* [19] proposed an interactive system for building fair rankings, and Xie *et al.* [17] introduced a visual system for fairness comparing rankings produced from graph mining recommender algorithms.

Evaluation of Fairness-Oriented Toolkits

Several researchers assessed toolkits that incorporate fairness into their process. Mashhadi *et al.* [21] studied the impact of the visualization styles of six open-source fair classification toolkits on student learning of fairness criteria. Lee *et al.* [22] evaluated the capabilities of open-source fairness toolkits and their suitability for commercial use through practitioner interviews and surveys. They found that many toolkits that contained visual representations of fairness were difficult for non-technical users to understand, even in tools like the What-If Tool [13], which were designed for broader audiences. Richardson *et al.* [23] conducted interviews

with machine learning practitioners to create a rubric for evaluating fairness toolkits. While there has been a surge in the development of fairness toolkits, Deng *et al.* [48] have highlighted gaps between fairness toolkits' capabilities and practitioners' needs.

Evaluation on Presentation of Fairness Information

Studies have evaluated the presentation of fairness related information in different scenarios. Van Berkel *et al.* [27] compared the perceived fairness level between text and scatterplot visualization techniques. The study found that the scatterplot visualization technique resulted in a lower fairness perception than text. Saxena *et al.* [49] investigated people's attitudes towards algorithmic definitions of fairness and found that people considered calibrated models, such as ratios, fairer than equal or meritocratic distributions in the context of loan decisions. Similar studies found that people perceive demographic parity and equalized odds as fair, depending on the scenario. Cheng *et al.* [50] compared three group fairness approaches in a child maltreatment predictive system. They found that people mostly supported equalized odds, followed by statistical parity and unawareness. Srivastava *et al.* [51] found that people prefer demographic parity among the 6 different notions of group fairness. Harrison *et al.* [52] conducted a user study on the perceived fairness of machine learning models in the criminal justice context and found conflicts between various inconsistent definitions of fairness. Nevertheless, Hannan *et al.* [53] showed that the factors of "what" and "who" matter in fairness perceptions and that the context of algorithmic fairness is more important in some domains than others.

Chapter 3

Visualization and Interaction Design

3.1 Goals and Activities Analysis

We characterize the data model and tasks for a decision-maker analyzing multiple stakeholder preferences and ultimately combining them to generate a fair consensus ranking.

We are given a set of **candidates**, described by **attributes**, to be ranked. One of the attributes, typically a categorical attribute referred to as the **protected attribute** (such as gender, race, or income-level), is associated with bias measurement and mitigation. We refer to candidates sharing the same value of the protected attribute as **groups**, such as Man, Woman, or Non-binary groups in the Gender attribute. Stakeholders in the committee (called **rankers**) each order (rank) the set of candidates to create a list of **base rankings** provided to the **decision-maker**. A decision-maker (head of the committee) using our system generates **consensus rankings** with the aim to order the candidates such that the base rankings, and thereby rankers, mostly agree with the consensus ranking.

The consensus ranking also must be fair. For auditing the fairness of rankings, we employ two metrics: a group-specific pairwise fairness metric **FPR** (Favored Pair

Representation) [6] to measure the fair treatment of each group in the ranking, and an aggregate fairness metric **ARP** (Attribute Rank Parity) [6] to quantify if the overall ranking across all groups satisfies the statistical parity fairness criteria [2]. In generating a fair consensus ranking, the decision-maker sets the **fairness threshold** value which controls the level of ARP represented in the consensus ranking. The later is then generated by a function utilizing the **Fair-Copeland Algorithm** [6]. A function **Kendall Tau distance** [54] computes the similarity/agreement between any two rankings.

We begin by employing task abstraction methodologies following procedures from Lam *et al.* [55] and recent works on group decision-making by Hindalong *et al.* [10, 11]. Table 3.1 outlines the goals and sub-goals for generating and analyzing fair consensus rankings that combines the preferences of multiple rankers (base rankings) into a single consensus ranking. For each sub-goal, we identified a set of visualization activities (Table 3.2) based on a widely used method in the visualization literature [56], leading to the design and implementation of several views.

3.2 FairFuse

We design FairFuse system to support the process of both analyzing and combining preferences from multiple stakehodlers into a *fair* consensus ranking. In designing FairFuse, we developed core views based on parallel coordinates, augmented with custom visual encodings for fairness metrics, and interactive components for generating fair consensus rankings. The system consists of several views and interaction capabilities to support the goals (Table 3.1) and activities (Table 3.2).

Table 3.1: Generic goals for rankings inspection and fair ranking generation and analysis

	GENERIC GOAL
G1	**Characterize Differences in Base Rankings**
	a Discover (dis)agreement on each candidate between rankings
	b Assess the discrepancy of candidates' position between base rankings
G2	**Investigate Protected Attribute**
	a Discover protected attribute groups of the candidates
	b Discover groups clustering of protected attribute in each ranking
G3	**Discover Bias in the Rankings**
	a Discover (dis)advantaged groups in each ranking
	b Investigate the treatment of groups across rankings
	c Intuit fairness of each ranking
G4	**Generate Fair Consensus Rankings**
	a Analyze multiple consensus rankings of different fairness level
G5	**Discover Nuances (not captured by the model)**
	a Analyze discrepancy on candidates between base rankings and consensus rankings
	b Re-evaluate Fair Consensus Rankings

3.2.1 Ranking Exploration View

The *Ranking Exploration View* uses parallel coordinates plot to explore and compare rankings of candidates between multiple stakeholders (A1, A2, A3, A16, A17) as shown in Figure 3.1E. It contains all candidates ordered into two or more base rankings. Columns on the left correspond to input base rankings; while fair consensus rankings generated by the decision maker are appended to the right upon their creation. Each candidate's set of attributes and values are represented by glyphs and colors [57] (A5, A6), collectively called a Candidate Card (Figure 3.2). By dragging-and-dropping the Candidate Card (A18), any generated consensus ranking can be adjusted if necessary.

The parallel coordinates is used to compare candidates across different rankings, drawing on features from similar rank oriented systems such as LineUp [7] and

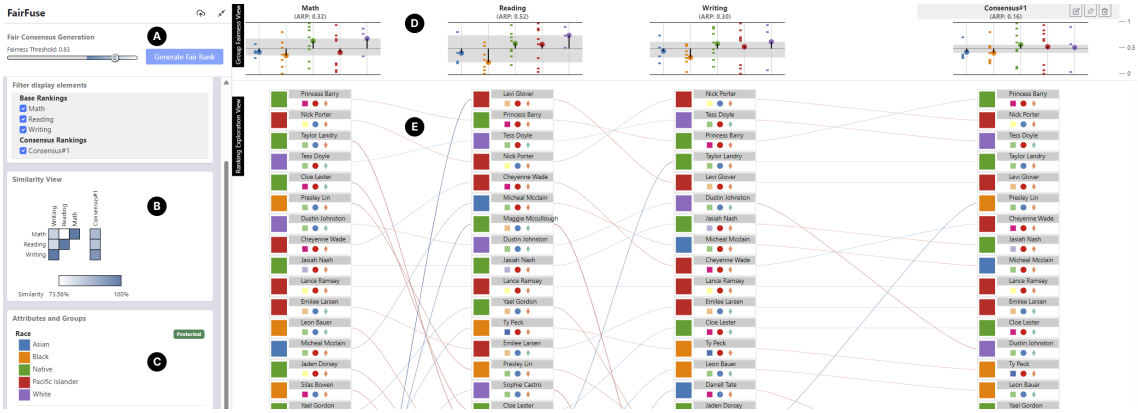


Figure 3.1: Design of the FairFuse system for constructing and analyzing fair consensus rankings. A) Consensus Generation, B) Similarity View C) Attributes Legend, D) Group Fairness View, and E) Ranking Exploration View

Hindalong *et al.* [10]. The order of candidates in a given column is based on candidate rank in the case of a base ranking columns, or the Fair-Copeland Algorithm in generated rankings. Each candidate appears across all rankings, with lines connecting them to illustrate change in position across rankings. Lines connecting the candidate across the rankings are colored based on the degree of change in the candidate's position between adjacent rankers. Candidates ranked higher in the subsequent ranking are colored in a gradient scale of blue, while those ranked lower are colored in a gradient scale of red.

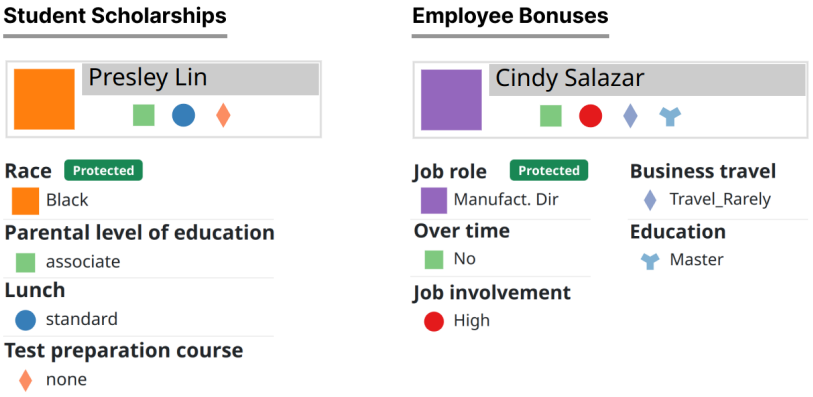


Figure 3.2: Candidate Card contains glyphs representing multi-variate attributes. The protected attribute is emphasized with a large (versus smaller) shape. Color represents the value of the attribute.

To reduce parallel coordinates clutter (*e.g.* [58]) while maintaining task effectiveness, we hide lines for which both candidates on adjacent rankers are not visible within the screen. Clutter can also result from orderings of parallel coordinate columns [59]. Users can drag to re-arrange columns, and FairFuse can be readily extended with automatic ordering techniques.

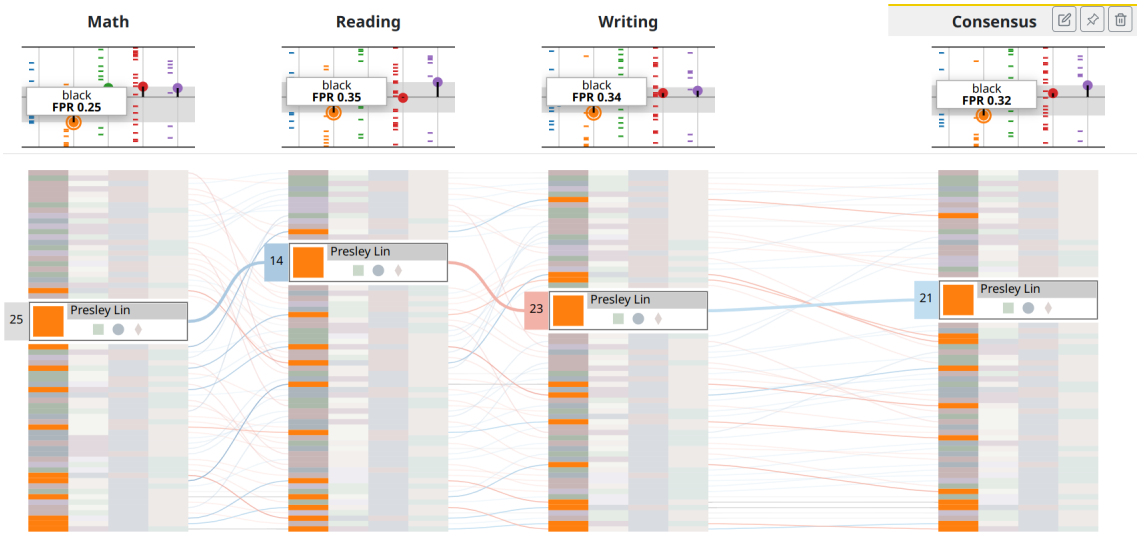


Figure 3.3: Compressed view with a group selected. When a group is selected in the Group Fairness View, all candidates of that group are highlighted, facilitating the decision-maker in focusing on both group and individual fairness considerations.

We also design a *Compressed Ranking View* mode (Figure 3.3) which represents a scaled-down version of the rankings. In this mode, the candidate cards (Figure 3.2) are initially hidden, but appear when hovering over a particular candidate. The protected attribute glyph is displayed with full saturation so that the decision-maker can explore how groups are distributed in each ranking, while other attributes are desaturated so as to be visible while interfering less with the protected attribute color.

3.2.2 Group Fairness View

The *Group Fairness View* (Figure 3.1E) captures fairness of a ranking at individual group level utilizing FPR score [6] (A6, A7, A8, A9, A10) and holistically across groups in the ranking using ARP score [6] (A11, A12, A14).

The FPR metric [6] captures if a specific group is fairly treated throughout the ranking). Specifically, FRP score = 0.5 denotes totally fair group treatment, while < 0.5 represents under-advantage and > 0.5 over-advantage. The ARP metric [6] captures if statistical parity fairness is satisfied by the ranking overall, *i.e.*, *all* groups are comparably treated to each other. Here, ARP = 0 is absolute fairness, anything higher is further and further from total fairness. This novel fairness view is critical to capture the notion that in multiple-group settings one or more groups may be fairly treated, while others may be unfairly over- or under-advantaged.

In designing the Group Fairness View, we initially explored 2 alternate prototypes (Figure 3.4). Because FPR and APR are scalar values, we first represented the FPR and ARP fairness scores with bar encodings at the top of ranking columns (Figure 3.4A). However, after determining that this design made it difficult to identify over-advantaged and under-advantaged groups, our second design placed an axis at FPR 0.5 and adopted a hybrid dot-plot and box-plot encoding (Figure 3.4B). This change supports more semantically meaningful visual queries. For example, dots above FPR 0.5 represent over-advantaged groups, informing the decision maker that they are unfairly receiving a larger share of favorable rank positions.

As explained below, our third and final design variation of the Group Fairness View as depicted in Figure 3.4C offers additional advantages. Since ARP measures the difference between the maximum and minimum FPR scores, we can visualize the ARP score with the region between the scores of the respective group. This change

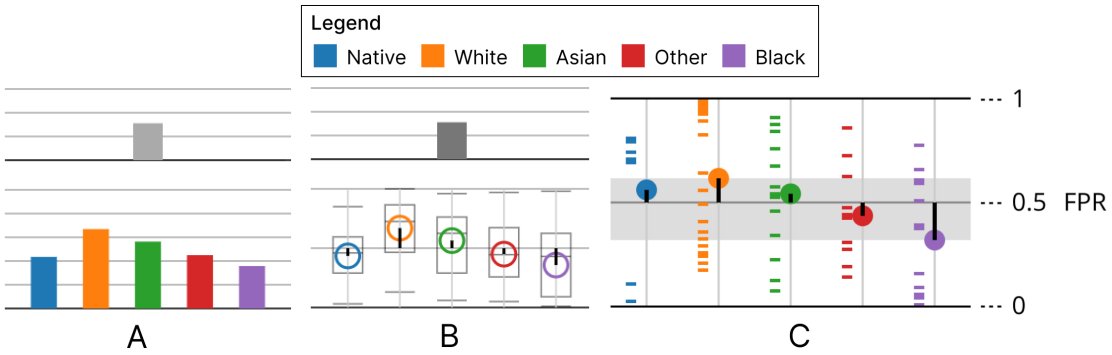


Figure 3.4: Group Fairness View design iterations. A) Colored barchart for FPR; gray bar for ARP. B) Dot plot for FPR, box-plot for distribution of groups in ranking. C) Final Design: Dot plot represents FPR, heatmap the distribution of groups, and shaded region the ARP.

enables visual queries within and across rankings to assess group fairness of each ranking, as a smaller ARP value would create a smaller shaded region. To mitigate the limitations of boxplots for showing non-contiguous distributions, we adopt a marginal mark-based distribution plot. Finally, the view affords interactive features, such as displaying exact FPR and ARP values on hover, and highlighting groups in the parallel coordinates plot on click.

3.2.3 Similarity View

The *Similarity View* (Figure 3.1B) uses a heatmap to show the similarity between any two rankings (A4, A15) with darker squares representing higher similarity between the rankings. This includes the ability to compare similarities between any two base rankings, and a base ranking with a consensus ranking. The similarity measure is calculated using a common measure for rank dissimilarity called Kendall-Tau distance [54]. With the help of tooltips and legend, the decision maker can extract the exact value of similarity between any two rankings.

3.2.4 Ranking Generation

The *Ranking Generation* process uses a button to first generate a consensus ranking without any fairness intervention. After the consensus ranking is displayed, the generation button is replaced with a slider (Figure 3.1A) – allowing the fairness threshold of generated consensus ranking to be adjusted (A13, A18). This process utilizes the Fair-Copeland algorithm [6]. Because a set of base rankings are unlikely to be completely unfair from the outset, the slider includes a gradient overlay to indicate that the fairness threshold will only produce fairer results if changed in a particular region. Similarly, on the other extreme, if the slider is set to 0, it will generate a consensus ranking solely based on the input base rankings.

3.2.5 Additional Interactions and Workflow

FairFuse provides additional interactions to support the decision-maker in a fairness-oriented rank analysis and generation workflow. Ranking Exploration, Group Fairness, and Consensus Generation and Similarity Views include design elements that respond to user actions such as clicks and hovers. A user hovering in the Ranking Exploration view, for example, will highlight a Candidate Card for easier exploration across views. A click in this scenario "pins" a candidate for comparison against other candidates. Brushing is also enabled [58], allowing the user to drag select ranges of candidates within particular columns, which is particularly useful in the compressed views. Similar hover and click functions are available in other views, mainly oriented towards emphasizing or de-emphasizing visual components to enable the decision-maker to focus on particular tasks.

To support iteration and adjustment of consensus rankings, FairFuse provides the decision-maker with editing features on consensus rankings. FairFuse supports manual

editing of fair consensus rankings, as the decision-maker may have additional context and information that they need to preserve in the resulting ranking. Decision-makers may adjust the fairness threshold of a consensus ranking to obtain another result, create or "pin" rankings, and manually adjust the position of candidates. Importantly, repositioning candidates immediately triggers the recalculation of fairness metrics, showing the decision-maker how fairness is lost or gained through their manual editing.

3.3 Use Case Scenarios Using the FairFuse System

A scholarship administrator, Jo, is responsible for determining the merit scholarship package of prospective students[1]. Jo needs to combine the recommendations of three rankers, teachers in Math, Reading, and Writing, and form a single ranking to allocate the merit scholarships. Cognizant that systemic and societal biases can affect how students of differing races perform in academic subject exams, which in turn can affect how students are perceived by subject-specific rankers, Jo seeks to detect and mitigate excess bias in the consensus ranking to ensure all groups are comparably treated.

Jo loads the data of base rankings given by the teachers along with candidate attribute information into the FairFuse. Jo uses the *Similarity View* to assess to what degree each base ranking agrees with others, along with visually inspecting the lines between adjacent rankings in the *Rank Exploration View*. At this point Jo uncovers that the Math teacher disagrees to some extent with the other teachers, which can make a consensus challenging, even before considering fairness.

Next, Jo switches to the compressed view to evaluate how candidate attributes are distributed across rankings. Auditing primarily for fairness, however, Jo pays

[1]An additional usage scenario is presented in supplemental material.

particular attention to the protected attribute, race. For this task, Jo studies the *Group Fairness View* on top of each ranking (Figure 3.1D), which shows distributions of protected attributes throughout the ranking. Jo notices immediately that the FPR fairness metric indicates that white students have a stronger advantage over students from other races. On closer examination, Jo discovers that across all rankings, students from the white group are clustered at the top, while students from the black group are clustered more towards the bottom. This is then reflected in the ARP scores (gray area) of the rankings, indicating the base rankings in general are far from fair as defined by statistical parity.

After exploring and comparing the similarities and fairness of the base rankings (Figure 3.1 B, D), Jo initiates the auto-generation of a consensus ranking, using the Consensus Generation. Immediately, Jo notices that the consensus ranking reflects the biases found in base rankings. Jo then progressively adjusts the Fairness Threshold (Figure 3.1A) to generate a fairer consensus ranking. Throughout this process, Jo references the Similarity View matrix and base rankings themselves to evaluate the extent to which base rankings are represented in the fair consensus. Honing in on a consensus ranking that balances the desired trade-off between the fairness and preference representation, Jo makes manual swaps between candidates to refine the target consensus ranking. With each edit, Jo's changes are audited visually by changes in the Group Fairness View (Figure 3.1D), helping ensure this manual manipulation does not drastically change the desired fairness measure. The resulting consensus ranking is both fair with respect to mitigating the over-advantage of white students and their disproportionately large merit awards, while ensuring the teacher recommendations expressed by base rankings are adequately combined and represented.

Table 3.2: Activities resulting from the goals and activities analysis, designed to support the goals in Table 3.1

	ACTIVITY
G1a	**Discover (dis)agreement on each candidate between rankings**
	A1 Locate each candidate across the rankings
	A2 Compare position of candidates across rankings
G1b	**Assess the discrepancy of candidates' position between base rankings**
	A3 Compare position of multiple candidates between rankings
	A4 Compare Kendall Tau distance [54] between rankings
G2a	**Discover protected attribute groups of the candidates**
	A5 Identify protected attributes of candidates
G2b	**Discover groups clustering of protected attributes in each ranking**
	A6 Locate candidates of each group in a ranking
	A7 Analyze distribution of candidates of each group
G3a	**Discover (dis)advantaged groups in each ranking**
	A8 Identify FPR score of each group
	A9 Compare FPR score with a baseline fair score
G3b	**Investigate the treatment of groups across rankings**
	A10 Compare FPR score of groups across rankings
G3c	**Intuit fairness of each ranking**
	A11 Identify ARP scores of the rankings
	A12 Compare ARP across rankings
G4a	**Analyze multiple consensus rankings of different fairness level**
	A13 Generate consensus rankings with different ARP thresholds
	A14 Compare ARP and FPR scores between rankings (including consensus rankings)
	A15 Compare Kendall Tau distance between rankings (including consensus rankings)
G5a	**Analyze discrepancies on candidates between base rankings and consensus rankings**
	A16 Compare individual candidate positions in base rankings with consensus rankings
	A17 Identify candidates with major differences in base rankings with consensus rankings
G5b	**Re-evaluate Fair Consensus Rankings**
	A18 Manipulate candidate position or Re-iterate fair consensus ranking generation with different fairness threshold

Chapter 4

Study Design

We now aim to investigate the challenges and opportunities of a system like FairFuse for the activities associated with fairness-oriented tasks. We adapt FairFuse into a new system, ConsensusFuse, by removing visualized fairness metrics and algorithms. ConsensusFuse system acts as a baseline for comparison in our study. Changes included the removal of 1) encodings of fairness metrics in the Group Fairness View (Figure 4.1H), 2) the Similarity View which uses metrics to compare the similarity of fair rankings, and 3) the fairness algorithm in the consensus ranking generation process, which had a slider to control the ARP [6]. Differences are shown in Figure 4.1.

In our study, we presented a scenario where participants were tasked with constructing a fair consensus ranking for scholarship distribution based on teachers' rankings of students. We performed a between-subjects study in which each participant was assigned to use either FairFuse or ConsensusFuse system.

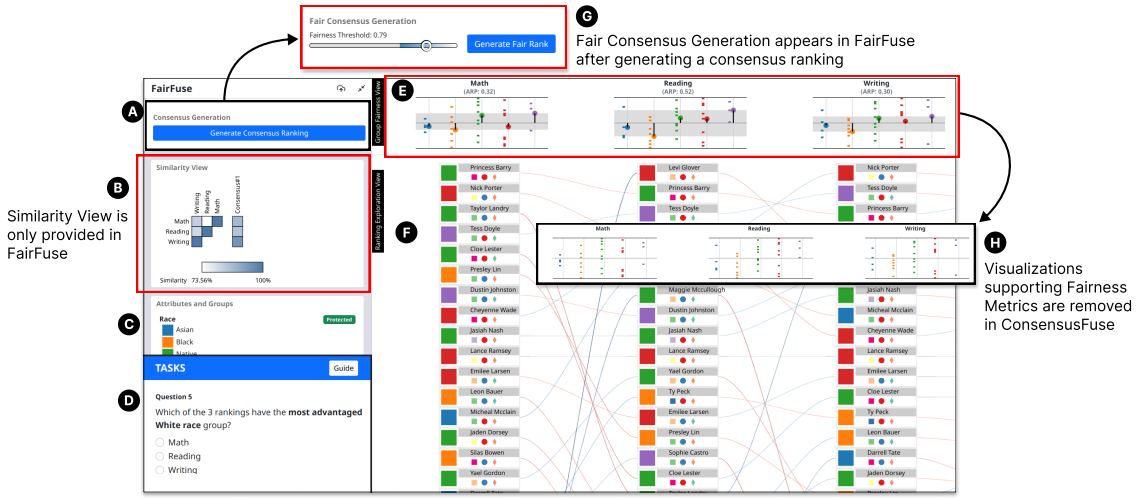


Figure 4.1: FairFuse and ConsensusFuse System Designs with changes in visualizations related to fair consensus generation. A) Consensus Generation, B) Similarity View (in FairFuse), C) Attributes Legend, D) Tasks presented to the participants, E) Group Fairness View (in FairFuse), F) Ranking Exploration View, G) Fair Consensus Generation (in FairFuse), H) Group View (in ConsensusFuse).

4.1 Procedure

We recruited 200 English-speaking participants agreeing to an IRB-approved consent form on Prolific, a crowd-sourcing platform. Based on multiple pilot studies, each participant was paid $5 USD for an estimated 25-minute study time, with an hourly rate of $12.00 USD. Since both the system used for the study is built for large screens, participants were filtered to use only desktop devices using Prolific's screening process. Our study consists of 3 phases: training, study, and feedback phase (Figure 4.2) as seen in a similar user study in the literature [60].

Training Phase. The study starts with the training phase (Figure 4.2A) introducing participants to different parts of the system through textual, figurative, and video explanations while also encourging them with analysis regarding consensus finding and bias mitigation.

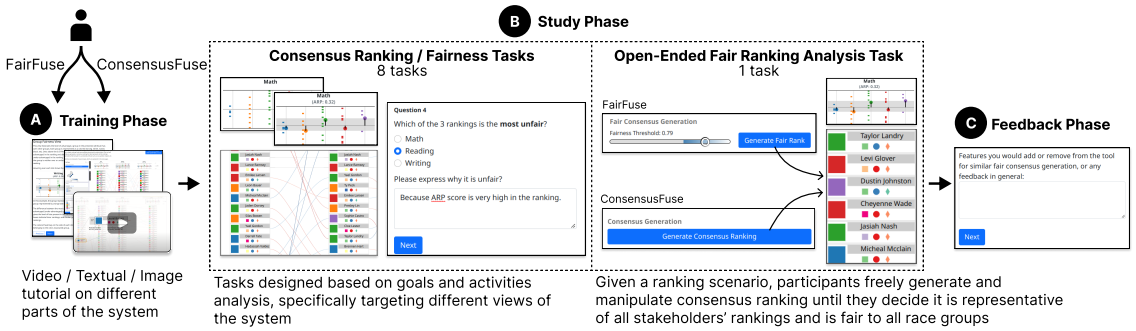


Figure 4.2: Study Design: We explore using visualization-enabled fairness metrics in building a fair consensus ranking. Participants are divided into two conditions, FairFuse: system with visualization-enabled fairness metrics and ConsensusFuse: system without visualizations for fairness metrics. Participants go through three phases. A) Training Phase, B) Study Phase and C) General Feedback Phase.

Study Phase. The second phase (Figure 4.2B) involved participants completing tasks. Both FairFuse and ConsensusFuse systems' interfaces were adjusted to include a view displaying the tasks. The sidebar was shortened to accommodate the tasks and participant answers at the bottom. During this phase, the participants interacted with the visualizations to find the answer(s). Each task was followed by a multiple-choice form with a dropdown or checkbox, and some were also followed by a free text form. The tasks in this phase were designed to increase in complexity gradually. Participants could refer back to the tutorial if they encountered difficulty. This phase was further divided into two parts. The first part focused on the systems' specific views and activities (Table 3.2) while additionally serving as a guided tutorial for the second part of this phase. On the other hand, the second part invited participants to interact with all system views while completing an open-ended task of constructing a fair consensus ranking.

Feedback Phase. The final phase of the study (Figure 4.2C) collected qualitative feedback on the system regarding generating a fair consensus ranking and demographics-related information.

4.2 Tasks Scenario Data

For this study, we adapted the data from the publicly available dataset [61] of students' rankings. The dataset contains multiple attributes, but for generating a consensus ranking, we used the relative ordering of students in three subjects, math, reading, and writing, as base rankings. Since our study phase has two parts, we created two datasets of 30 students each, where one dataset was used for each of the two study parts. The dataset was split such that both had all 5 groups of the protected attribute, race, the advantaged and disadvantaged groups can be separable. Race was the protected attribute for both datasets, with five groups: White, Native, Black, Asian, and Pacific Islander.

4.3 Study Task Design

Targeting the goals and activities (Table 3.2), we created a set of tasks for the participants, listed in Table 4.1. The first eight tasks focus on different individual views of the system. These tasks encompass the Ranking Exploration View with candidate cards containing attributes of the candidate and parallel coordinates plot of the rankings, Similarity View, Group Fairness View, and the Consensus Generation process. The final task asks participants to conduct a free-form fair consensus ranking generation.

Table 4.1: List of task prompts given to the participants. Tasks are targeted at the *Goals and Activities* (Table 3.2) analysis.

	Task / Task Prompt	**Target Activity**
T1	**Locating protected attribute** What is the race of Taylor Landry?	A1, A5
T2	**Identifying Advantaged Group(s)** Which race groups are over advantaged in the Math ranking?	A6, A7, A8, A9
T3	**Visualization Use** Click on the visualization you primarily used to deduce the answer for the previous question?	
T4	**Identifying Attribute-level Unfairness** Which of the 3 rankings is the most unfair? Please express why it is unfair?	A11, A12
T5	**Identifying Group-level Unfairness** Which of the 3 rankings have the most advantaged White race group?	A6, A7, A10
T6	**Utilizing PCP Position Comparison** How is Taylor Landry's position ranked in Math compared to Reading?	A2, A3
T7	**Interpreting PCP Gradient** Select the candidate with most disagreement between Math and Reading rankings. Please explain how you deduced your answer.	A2
T8	**Using Consensus Generation Procedure** STEP 1: Generate a consensus ranking using the button on the top of the left sidebar. STEP 2: Use the pin icon in the heading of the generated ranking to pin the ranking. STEP 3: Please identify which base ranking is most dissimilar to consensus ranking you just generated.	A4, A13, A15
T9	**Using Fair Consensus Generation Procedure** Generate a fair consensus ranking that: 1. Is representative of all the base rankings 2. Does not over or under advantage race groups	A1 - A18

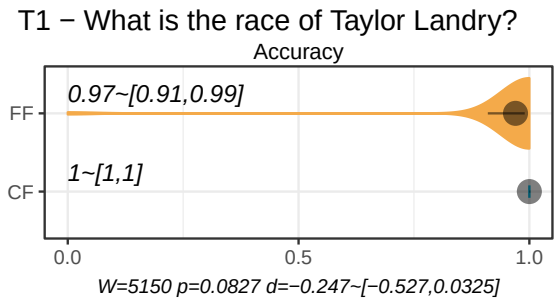
Chapter 5

Results

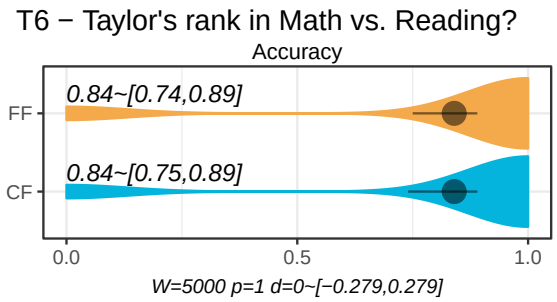
We recruited 200 participants (a number obtained via power analyses following pilot studies) and evenly divided them into two groups, namely, FairFuse and ConsensusFuse. We computed 95% confidence interval using a bootstrapped method and effect size using Cohen's *d*. Our results also include p-value (p) from the Wilcox Test (W).

5.1 Ranking Exploration Tasks

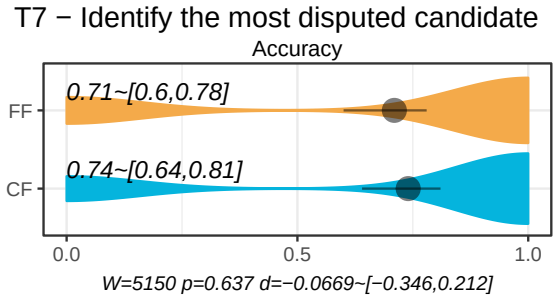
Since three of the tasks, **T1**, **T6** and **T7**, relied on the unmodified views presented for both groups, we observe that there is no significant difference in the answers given by the participants. The violin plot with confidence intervals, p-value and effect size are shown in Figure 5.1. We report no significant difference in all three tasks between the two conditions, namely, $p = 0.0827$, $p = 1.0$, and $p = 0.637$, respectively. We find that participants are able to identify attributes and compare positions of candidates between rankings using Parallel Coordinates Plot in both systems. For **T7**, which is an advanced task compared to **T1** and **T6**, we see a slight decrease in the correct answers. **T7** asked participants to identify the candidate with the most



(a) Locating protected attribute [T1]



(b) Utilizing PCP Position Comparison [T6]



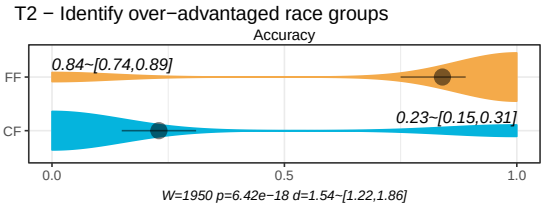
(c) Interpreting PCP Gradient [T7]

Figure 5.1: Results for Ranking Exploration

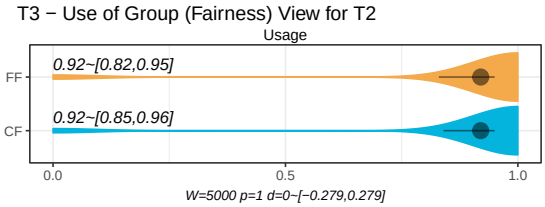
disagreement between two rankings. This task involved identifying a candidate card connected with a line between two adjacent rankings with the most inclination.

5.2 Fairness-oriented Tasks

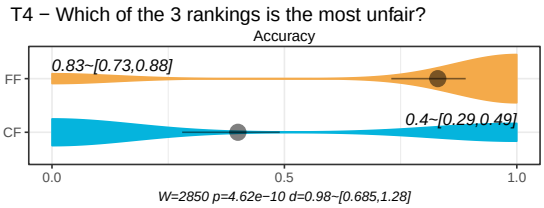
T2 asks participants to identify advantaged groups in one of the three rankings provided. During the experiment, participants were provided with checkboxes of five race groups allowing them to select multiple race groups. The ground truth included two advantaged race groups based on the FPR scores. We observe that the user performance in FairFuse ($M = 0.84 \sim [0.74, 0.89]$) is significantly better than ConsensusFuse ($M = 0.23 \sim [0.15, 0.31]$) as shown in the violin plot (Figure 5.2a) with a large effect size ($d = 1.54 \sim [1.22, 1.86]$). The careful design of the Group



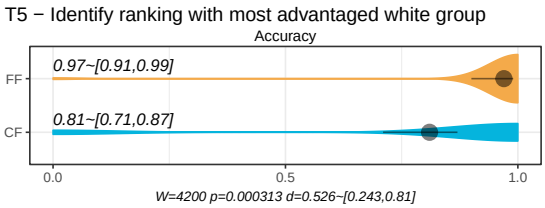
(a) Results for Identifying advantaged Group(s) [T2] (both native and white as correct answer)



(b) Results for use of Group (Fairness) View for T2 [T3]



(c) Results for Identifying Attribute-level Unfairness [T4]



(d) Results for Identifying Group-level Unfairness [T5]

Figure 5.2: Results for Fairness Intuition

Fairness View in FairFuse with the affordance of a horizontal line providing a visual cue of the baseline that separates the advantaged from disadvantaged groups could have helped FairFuse achieve better accuracy for this question. We also find that both FairFuse and ConsensusFuse participants use the same view for tackling this question **T2** as seen in Figure 5.2b. It's noteworthy that the majority of participants in the ConsensusFuse study selected one of the two correct advantaged groups, while the participants in FairFuse identified both correct advantaged groups (as shown in Figures 5.3a and 5.3b). This highlights the significance of fairness metrics and visualizations in identifying multiple advantaged or disadvantaged groups when a large number of groups are involved.

For **T4**, participants were asked to identify the most unfair ranking among the three rankings provided. While **T2** focused on the level of advantage each group has using FPR measure [6], **T4** focused on utilizing the ARP metric [6]. Similar to **T2**,

with **T4**, we get significantly different results between the two conditions with high accuracy in FairFuse (ConsensusFuse: M = 0.4 [0.29, 0.49] vs. FairFuse: M = 0.83 [0.73, 0.88]) as shown in Figure 5.2c. We also instructed participants to express why they think their ranking choice is unfair. Two types of reasoning are found most commonly in the FairFuse condition: expression at the vis-level and expression at the understanding level. Expression at vis-level reports the ARP score or visualization that mimics the ARP score, such as:

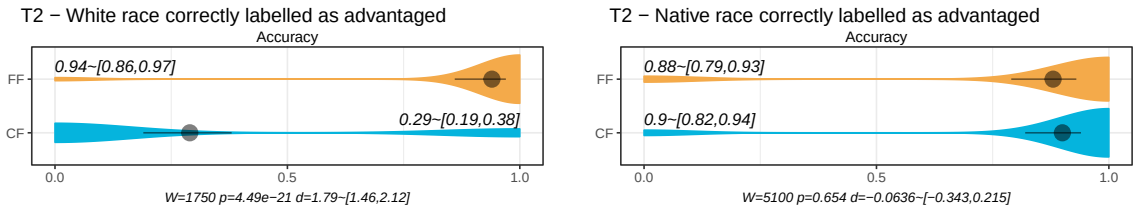
> *The grey bar is the widest with [the] highest ARP index.*

Expression at the understanding level goes beyond just reporting the ARP score, such as:

> *Reading shows the largest disparity between the highest and lowest group fairness scores, ergo the disparity between highs and lows would be the most unfair.*

ConsensusFuse participants tend to consider only a single group resulting in incorrect answers, such as:

> *The black group is very under-advantaged and is ranked a lot lower than other groups.*



(a) Results for Identifying Advantaged Group(s) with **white** as correct answer [T2]

(b) Results for Identifying Advantaged Group(s) with **native** as correct answer [T2]

Figure 5.3: Results per individual group for Identifying advantaged Group(s) [T2]

Also, it is interesting that some ConsensusFuse participants did meticulous calculations of individual groups, such as:

> *100% of the white students are in the top half, but only 28.5% of the black students are.*

Task **T5** builds from **T2** and **T4**, where participants were asked to identify the ranking with the most advantaged White race group. We find a small but significant difference in accuracy ($p = 0.000313$; FairFuse: $M = 0.97 \sim [0.91, 0.99]$ vs. ConsensusFuse: $M = 0.81 \sim [0.71, 0.87]$) with medium effect size ($d = 0.526 \sim [0.243, 0.81]$) as shown in Figure 5.2d. This may be because **T5** specifically asks about a particular group instead of multiple groups resulting in similar results like Identifying Advantaged Group(s) (**T2**) with native as a correct answer (Figure 5.3b).

5.3 Consensus Representation and Analysis Tasks

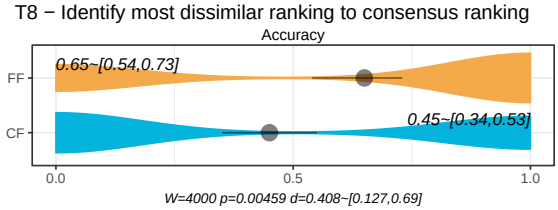


Figure 5.4: Results on Using Consensus Generation Procedure [T8]

To assess FairFuse's performance in identifying similarity of consensus ranking to base rankings, we device **T8**. To ensure a fair comparison between the systems, we asked both groups to start with generating a fairness-unaware consensus ranking, followed by selecting the most dissimilar base ranking. This way, both groups have the same state of rankings to begin with. The violin plot shows the result (Figure 5.4) with a significant difference between the two groups and a medium effect size

($p = 0.00459$, $d = 0.408 \sim [0.127, 0.69]$). Although FairFuse ($M = 0.65 \sim [0.54, 0.73]$) was more accurate than ConsensusFuse ($M = 0.45 \sim [0.34, 0.53]$), the advantage was not very high.

For this particular task, we also asked participants to reason their choice of answer. We find that most of the participants in FairFuse, even though they correctly identify the most dissimilar ranking, seems to have mentioned the use of Group Fairness View instead of the Similarity View, such as:

> *The ARP of reading is the furthest away from the ARP of the consensus.*

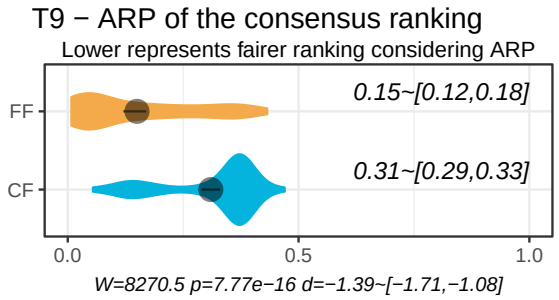
Interestingly, despite having the Similarity View in FairFuse, some participants either used a process similar to that of the ConsensusFuse participants by dragging individual base rankings towards the consensus ranking and counting line crossings, or didn't find the view useful.

> *I dragged each individual ranking over to place it side-by-side with the consensus ranking. [...] reading had the most lines that strayed from this path.*
>
> *I did not find the Similarity View very helpful.*

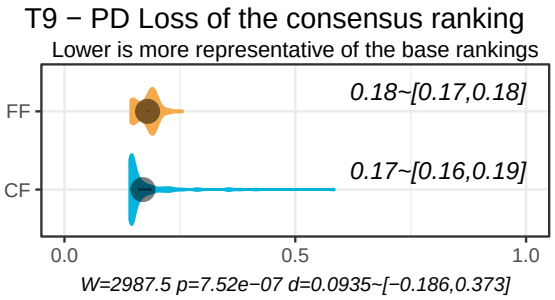
As a result, while the quantitative data supports the expectation that FairFuse would perform better in identifying the (dis)similarity between the consensus ranking and base rankings, with a statistically significant difference, the qualitative results do not fully support this conclusion. Participants focused more on the fairness metrics and disregarded other crucial information like the Similarity View.

5.4 Open-Ended Fair Ranking Analysis Task

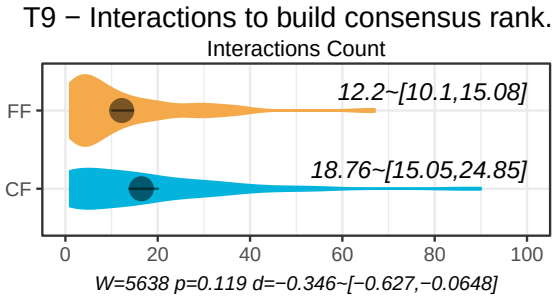
In **T9**, we ask participants to generate a fair consensus ranking that is representative of all the base rankings such that it does not over or under-advantage race groups.



(a) Results for ARP of generated fair consensus ranking [T9]



(b) Results for PD Loss of generated fair consensus ranking [T9]



(c) Total interactions the participants made to build fair consensus ranking [T9]

Figure 5.5: Results for Open-Ended Fair Ranking Analysis Task

We analyze the ARP scores between the two groups (which ranges from 0 to 1, with 0 representing a ranking with perfect statistical parity [6]) to measure the group fairness requirement. We find that FairFuse participants generally agree on consensus rankings with lower ARP scores ($M = 0.15 \sim [0.12, 0.18]$) compared to ConsensusFuse ($M = 0.31 \sim [0.29, 0.33]$) with a large effect size ($d = -1.39 \sim [-1.71, -1.08]$), interpreting that the participants fail to create a fairer consensus ranking in ConsensusFuse. However, we note that some of the participants, even without the fairness metrics and its visualizations, built consensus rankings with low ARP scores.

We observe that the PD Loss [6] (representation of base rankings in the consensus

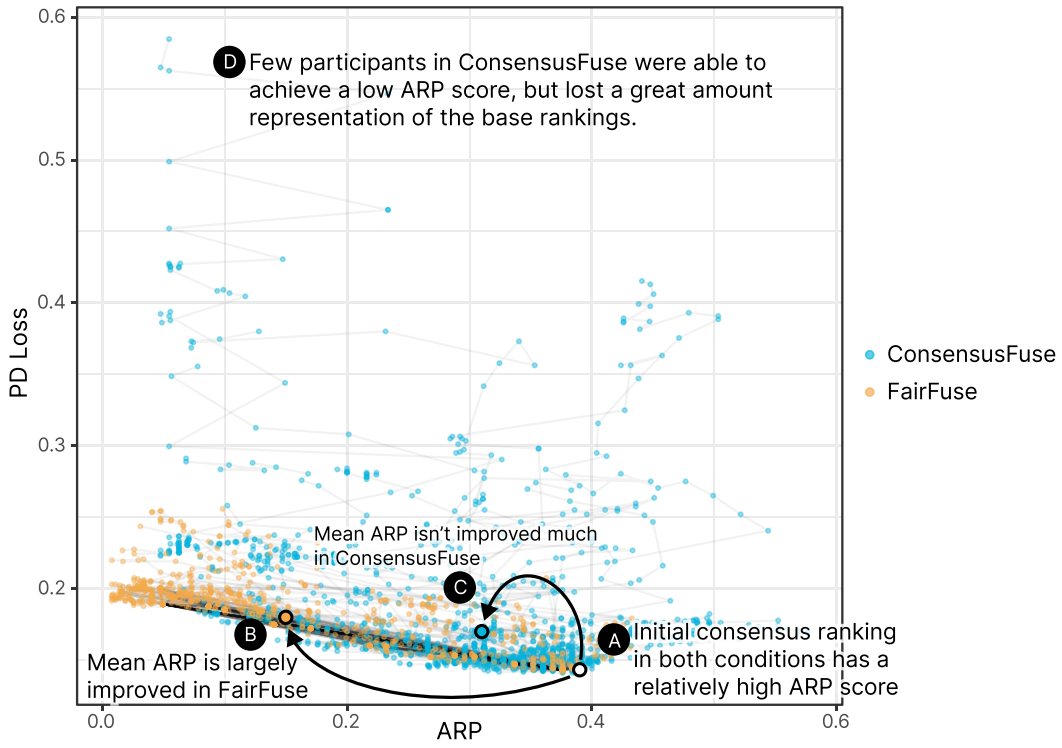


Figure 5.6: Results of ARP vs. PD Loss throughout each user interaction while generating a fair consensus ranking. The white dot indicates the ARP and PD Loss of the initial consensus ranking in both conditions.

ranking, with 0 representing that all the base rankings exactly match the consensus ranking) in both groups are similar (Figure 5.5b) despite some participants in ConsensusFuse ending up producing rankings that are far in distance from the base rankings, yet on the fair side, as seen in scatterplot (Figure 5.6D). Figure 5.6A marks the initial consensus ranking for both conditions, which has a relatively higher ARP score. Interactions included drag-and-drops of candidate cards for updating the consensus ranking and generation of consensus rankings. Figure 5.6B marks the vastly improved mean ARP value in FairFuse compared to Figure 5.6C in ConsensusFuse. We find that FairFuse participants make fewer interactions to agree on a fair consensus ranking as shown in the violin plot (Figure 5.5c) (ConsensusFuse: $M = 18.76 \sim [15.03, 24.62]$ vs. FairFuse: $M = 12.2 \sim [10.1, 15.08]$).

Chapter 6

Discussion

Overall results suggest that while both systems are suited for exploring ranking-related tasks, FairFuse outperforms in terms of accuracy in fairness-related tasks. Also, fewer interactions are involved in generating fair consensus rankings in FairFuse. We find that FairFuse, with its unique visualization-enabled fairness metrics, helps keep a balance between generating a ranking that maximizes the agreement of base rankings while keeping it as fair as possible concerning statistical parity, a common definition of fairness. However, we also find that users are drawn towards relying on fairness metrics and algorithms to complete the tasks, sometimes erroneously so. This introduces a tension between the goals of building a representative consensus versus ensuring that it is fair– a tension that creates interesting constraints and challenges for design.

Based on our study, we distilled a set of **4** takeaways summarizing how we observed visualized fairness metrics and algorithms helping or hindering in tasks and decision-making contexts. These takeaways may hold broader implications for developers of fairness metrics and algorithms, designers of visual interfaces, and the fairness community at large.

1) **Help: Researchers developing fairness-aware algorithms should incorporate ways for end-users to tune fairness, relative to other problem objectives.**

Given evolving societal norms and values, definitions of fairness can change over time and place. Definitions also vary from one discipline to another [62]. Algorithms designed to assist in incorporating fairness incorporate ways for decision-makers to tune fairness in the specific problem context. This increases both agency on the part of decision-makers, and incorporates their specific domain knowledge and worldviews. While FairFuse could produce an absolute fair consensus ranking based on the algorithm used, we find that participants set the fairness threshold close to the absolute threshold to generate a fair consensus ranking. This behavior suggests that allowing individuals to adjust the parameters of an algorithm can lead to more satisfactory and appropriate results. Moreover, making fair algorithms tunable allows for more transparency and accountability in decision-making, as decision-makers can see and understand the factors influencing the algorithm's output.

2) **Hinder: Visualization designers should be mindful that visually displaying fairness metric may lead to increased credence in and over-reliance on metrics**

Our results suggest that decision makers' decisions align with visualization-enabled fairness metrics (Figure 5.6). From a positive perspective, alignment with fairness metrics can promote fairness in decision-making. Yet, designers should also be cautious about the consequences of such drift. Nudging decision-makers toward visual indications of fairness may result in decision-makers blindly trusting such metrics and algorithms could miss the societal nuances

that the metrics cannot capture, which is reflected in participants' comments, *e.g.*: *"Fairness threshold [of] 1 seems to do the job?"*, *"I use the slider and slide it to fairness threshold to 1. [...] Then the ranking will be unbiased."* Visualization designers and the fairness community should be mindful of the potential for "fairness drift", particularly as metrics are increasingly incorporated into visual interfaces.

3) **Help: Properly designed visualizations of fairness metrics can help people navigate complexity in decision-making contexts**

The multi-objective nature of fairness related tasks can be tricky to navigate for non-expert users where achieving a goal (such as a building a good consensus ranking) is also subjected to bias mitigation. Inclusion of large number variables can make it worse as we see in our results where participants were able to identify only one of the two advantaged groups without the help of visualizations supporting fairness metrics (Figure 5.3). Identifying such groups can highlight areas of concern, making it easy for further analysis in mitigating bias. Properly designed visualization of fairness metrics can help identify bias across a larger number of variables helping individuals to make informed decisions in the decision-making process.

4) **Hinder: Improperly designed fairness metrics visualizations can lead people to incorrect conclusions**

While visualization tools like FairFuse can be used to promote fairness in building a consensus ranking, it is crucial for visualization designers to be mindful of the way in which fairness metrics are presented, as improper design can lead individuals to draw incorrect conclusions. For example, in the case of FairFuse, presenting new visualization such as the Group Fairness View

on occasion led participants to overlook other important information such as the Similarity View (see Section 5.3), Yet, the later is equally important in maintaining consensus. Failure to do so could result in an incomplete understanding of the task at hand.

Chapter 7

Limitations and Future Work

Both the FairFuse system [63] and the crowd-sourced study comes with certain limitations. FairFuse focusses on ARP and FPR fairness metrics [6] within the widely accepted definition of group fairness in the fairness community. It also considers one tunable algorithm for generating fair consensus ranking. The fulfillment of the goals of the system relied on those metrics in the tasks abstraction phase. However, Verma and Rubin [64] highlight that a decision considered fair by one definition may be deemed unfair by others, and laypeople's judgment often aligns with simple notions of fairness like group fairness [65]. Therefore, future work could incorporate multiple fairness definitions and algorithms, and conduct similar user studies. Future studies might also examine the potential benefits and drawbacks of using tunable algorithms like in FairFuse for fairness-related tasks. In addition, these studies could assess the impact on decision-makers trust in these systems and the possibility of an increased cognitive load.

Chapter 8

Conclusion

The concern for fairness in AI tools and online platforms has amplified the need for effective methods of identifying and mitigating bias in ranking processes. However, the complexities of fair consensus ranking, including multiple bias-causing factors and nuanced ethical and societal values, make a fully automated system unreliable. Human-in-the-loop systems, which offer a comprehensive approach to bias mitigation, can be valuable, but there is limited evidence on the benefits and challenges of designing visualizations that support fairness metrics.

To investigate these challenges, we built a system for building fair consensus rankings, FairFuse. We then conducted a crowd-sourced study across goals and activities designed for building a fair consensus ranking between a metrics-based visualization FairFuse and a non-metric based visualization system ConsensusFuse. Our findings suggest that well-designed visualizations can aid in creating fair consensus rankings, but they may also hinder certain tasks, particularly balancing goals beyond fairness in decision-making contexts.

Appendix A

Appendix

A.1 Datasets

For the demonstration of generating fair consensus rankings, we used the following publicly available datasets for two different scenarios:

1. Scholarships Distribution scenario: We used the dataset of 60 students with students scores in 3 subjects: Maths, Reading and Writing. The dataset contains multiple attributes but we utilize the exam scores in the dataset to convert it into 3 base rankings and build a fair consensus rankings by taking race as a protected attribute. Within the race attribute, there are 5 groups. `http://roycekimmons.com/tools/generated_data/exams`

2. Employee Bonus Distribution scenario: We used a subset of dataset of Employee Attrition and Performance. We utilized the performance rating in the dataset to convert it into 3 base rankings (R1, R2 and R3). We randomized the rankings between these rankings so that they resemble the disagreement in rankings as in real-world. `https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset`

Name	Symbol	Representation
Candidates	$X = x_1, ..., x_n$	n candidates ranked
Base rankings	$R = r_1, \ldots, r_m$	m rankings produced by multiple stakeholders (rankers)
Protected Attribute	p	Attribute for bias measurement/mitigation such as gender
Favored Pair Representation [6]	$FPR(G_{p:v}, r)$	Measure quantifying if a group $G_{p:v}$ is fairly treated in ranking r. Ranges from $[0, 1]$ where value 0.5 denotes fair treatment, 0 denotes complete disadvantage, and 1 denotes over-advantage.
Attribute Rank Parity [6]	$ARP(p, r)$	Measure quantifying if ranking r is fair with respect to statistical parity for p. Ranges from $[0, 1]$, 0 denotes absolute fairness, and 1 denotes complete unfairness.
Kendall-Tau [54]	$distKT(r1, r2)$	Ranking similarity metric comparing rankings r_1, and r_2. The value represents the number of pairs swap needed to convert r_1 to r_2.
Consensus Ranking	C	Ranking that is closest to the base rankings, thereby best representing R.
Fairness Threshold	t	Parameter controlling the maximum ARP of the fair consensus ranking, i.e., maximum resulting ARP of fair consensus ranking $= 1 - t$
Fair-Copeland Algorithm [6]	$FC(R, X, p, t)$	Method to generate a fair consensus ranking, such that all preferences in R are maximally represented subject to the result having at most ARP of $1 - t$

Table A.1: Data model used for fair consensus ranking generation

These datasets do not contain individual names. So, we used the names python package [1] for generating names.

A.2 Detail of Data Model

A.2.1 Input

Our system FairFuse is designed to aid consensus decision-making in regards to a set of candidates for the task at hand (e.g., candidates for scholarships). We assume

[1] https://pypi.org/project/names/

that in the set of n candidates X, each candidate is described by a set of attributes $\mathcal{A}$, and categorical **protected attribute** p such as gender, race, or income-level. We refer to p as the protected attribute chosen by the user for bias measurement and mitigation. For each possible value v of the protected attribute p, there is a **group** $G_{p:v}$ composed of individuals in set X that have the same value v for the protected attribute p. For instance, $G_{race:asian}$ is the group of all candidates that have the value asian for the race protected attribute.

The decision-maker utilizes our system to explore and combine (into a consensus decision) the preferences over X of multiple stakeholders. We refer to these stakeholders as as **rankers**. Each of the m rankers provides a ranking r_i ordering the candidates in X. We collectively refer to the rankings produced by the rankers as a set of **base rankings**, $R = r_1, ..., r_m$.

A.2.2 Metrics

In quantifying bias in the base rankings R our system employs the group fairness notion of **statistical parity** [2]. Statistical parity is a contemporary fairness notion stipulating that candidates must receive an equal proportion of the positive outcome regardless of their protected attribute value. In our setting, the positive outcome is favorable rank positions. Thus, when assessing statistical parity we measure if in the given ranking (a base ranking or consensus ranking) groups have equal proportions of favorable rank positions. To quantify statistical parity fairness we use the metrics introduced in Cachel et al. [6]. Namely, we utilize the metrics of FPR (Favored Pair Representation) and ARP (Attribute Rank Parity). For auditing rankings, the former quantifies how fairly a specified group is treated, while the later quantifies the presence (or lack) of statistical parity.

ARP and FPR are pairwise metrics. Any ranking r over n candidates can be

decomposed into (n(n - 1))/2 pairs of candidates (x_i, x_j) where $x_i \prec_r x_j$. Specifically, the metrics count mixed pairs, where a mixed pair is pair comparing candidates from two different groups. For instance, a pair with two woman candidates is not a mixed pair, while a pair with a man and woman is a mixed pair. Intuitively, the more mixed pairs a group is favored in (or "wins"), the higher up in the ranking that group is compared to other groups.

The FPR metric ranges from $[0, 1]$, where a value of 0.5 indicates the given group in the specified ranking is fairly treated [6]. A value of 0 indicates the group is totally disadvantaged (i.e, occupying the bottom and thus worst rank positions), and a value of 1 indicates the group is over advantaged (i.e, occupying the top and thus best rank positions). Equation A.1 shows the calculation for the FPR measure for a given group $G_{p:v}$ in ranking r, where $\omega_M(G_{p:v}, r)$ is the total number of mixed pairs in ranking r.

$$FPR(G_{p:v}, r) = \sum_{x_i \in G_{p:v}} \sum_{x_l \notin G_{p:v}} \frac{countpairs(x_i \prec x_l)}{\omega_M(G_{p:v}, r)} \tag{A.1}$$

Next, the ARP measure utilizes the FPR scores for all the groups in order to quantify if the given ranking satisfies statistical parity. Specifically the ARP measure is the maximum absolute difference between FPR scores. The ARP measure ranges from $[0, 1]$, when $ARP = 1$ then the protected attribute is maximally far from statistical parity and the ranking is maximally unfair [6]. Meaning, one group is entirely at the top of the ranking, while a second group is entirely at the bottom of the ranking. When $ARP = 0$, perfect statistical parity is achieved and the ranking is totally fair. Equation A.2 shows the calculation for the ARP measure for a given protected attribute p in ranking r.

$$ARP(p, r) = \underset{\forall\ (G_{p:v}, G_{p:j}) \in X}{\operatorname{argmax}} |FPR(G_{p:v}, r) - FPR(G_{p:j}, r)| \tag{A.2}$$

Finally, to measure how similar rankings are (both base and consensus rankings) we employ the rank similarity Kendall-Tau distance [54]. Equation A.3 expressed the Kendall-Tau distance between two rankings r_1 and r_2.

$$\begin{gathered} dist_{KT}(r_1, r_2) = \\ |\{\{x_i, x_j\} \in X : x_i \prec_{r_1} x_j \text{ and } x_j \prec_{r_2} x_i\}| \end{gathered} \tag{A.3}$$

A.3 Use Case Scenario for Employee Bonus Distribution

A manager is tasked with deciding annual bonuses for employees in their organization. They setup a committee with leads from different departments within the organization who rank all the employees. The manager now needs to combine each of the committee preferences into a final consensus ranking so that bonus awards can be allocated to employees. However, they also want it to be fair with respect to employees from different job roles to not advantage certain role at bonus time. To tackle this challenge, all the rankings made by the committee members are uploaded to the FairFuse tool with job roles as a protected attribute.

The manager immediately notices that rankings by the committee have a lot of disagreement by inspecting the number of crossings in the Ranking Exploration View. Looking at the Group Fairness View, they notice that the Human Resource (HR) group is highly advantaged while Research Directors are highly disadvantaged across all three rankings. Upon generating a consensus ranking, the manager notices this bias is transferred and reflected in consensus ranking as well. So, they decide to mitigate bias by adjusting the Fairness Threshold Slider and re-generating the

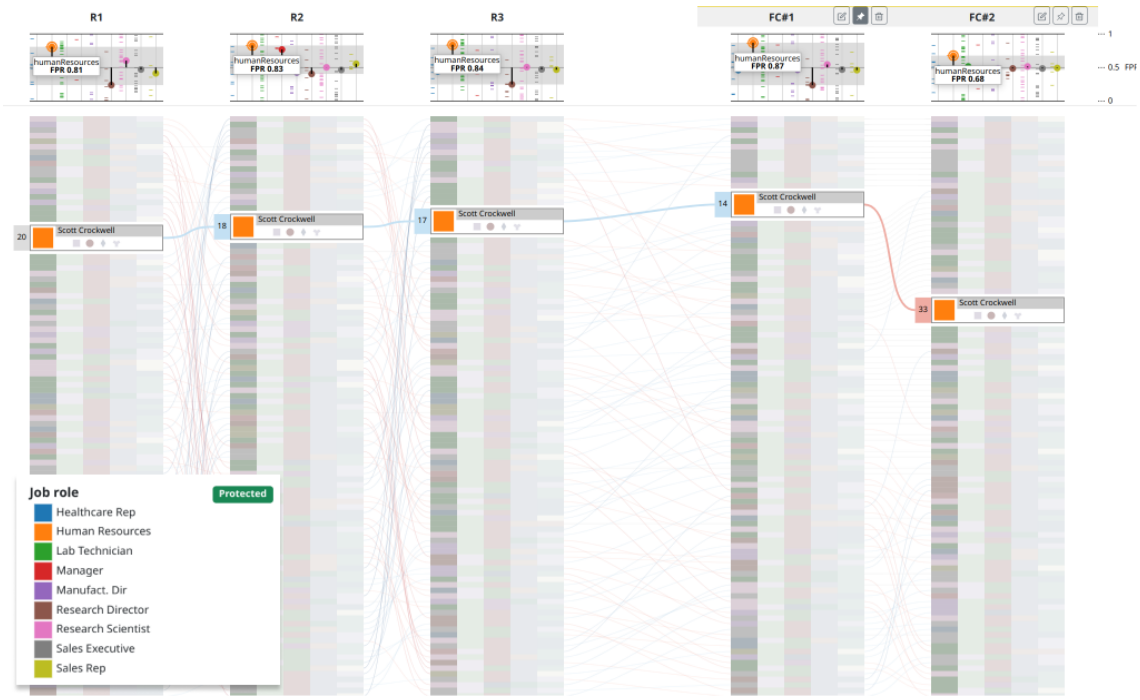


Figure A.1: Selecting a group in the Group Fairness View highlights candidates of that group. In this case, there is only single candidate representing Human Resource Group.

consensus ranking. While the ARP score is significantly reduced, they notice that HR group is still advantaged with higher FPR score.

Looking back at the heatmap in the Group Fairness View, there seems to be a single employee as HR. So, to inspect this employee they select the dot that represents HR. The HR employee is highlighted in the ranking exploration view as well. They click on the highlighted employee in the ranking exploration view and realize that the employee's rank does not change much between base rankings by the committee members, but the generated fair consensus ranking lowers this employees position (Figure A.1). The employer does not like this as the employee is the only one as an HR. Hence, they decide to manually adjust the ranking of the employee by dragging the employee card higher in the rank. Happy with the final fair consensus ranking, the manager proceeds with distributing bonuses based on the ranking.

A.4 Demographics of the crowd-sourced user study

Demographics of the participants pool.

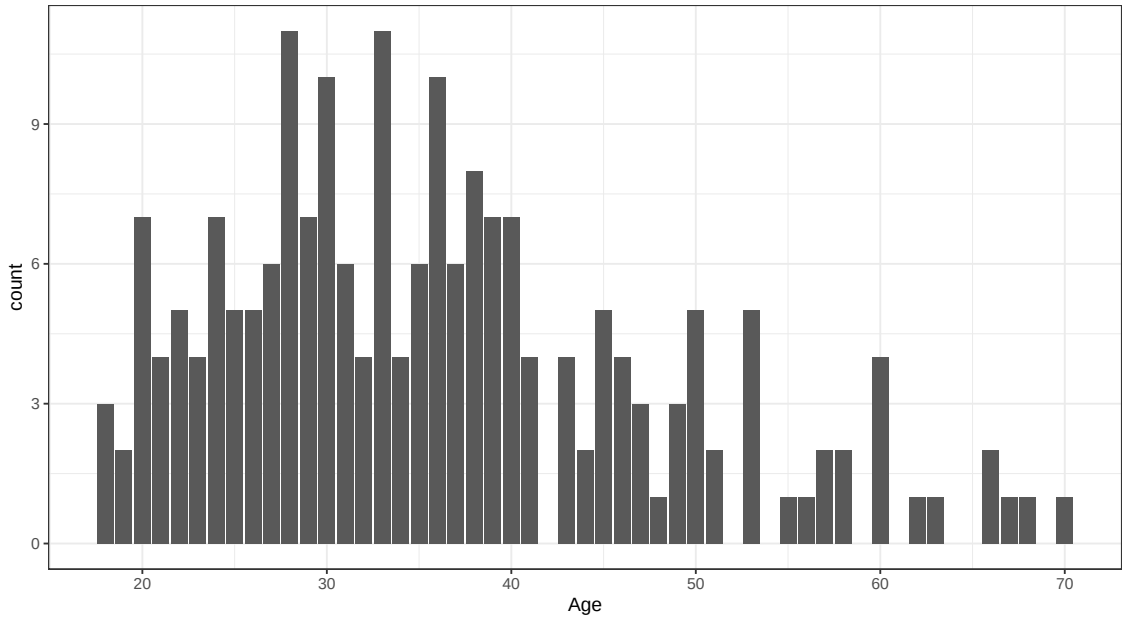


Figure A.2: Age Distribution

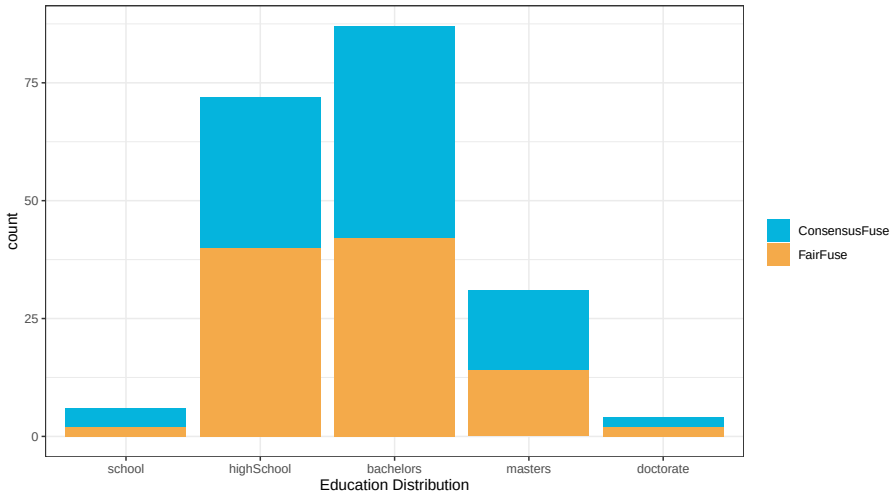


Figure A.3: Education Level Distribution

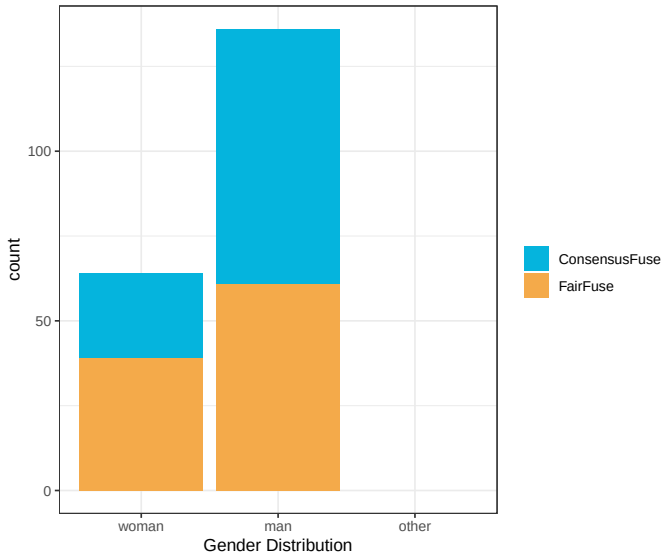


Figure A.4: Gender Distribution

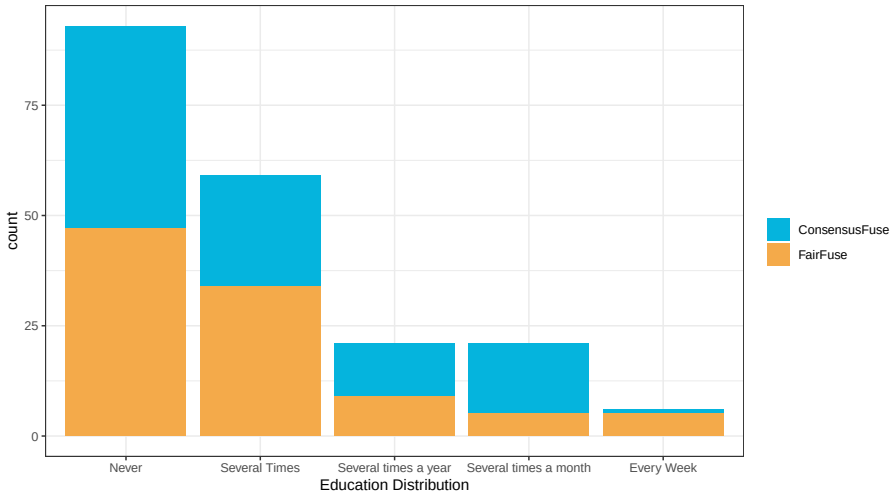


Figure A.5: Visualization Experience Distribution

Bibliography

[1] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[2] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proc. 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[3] John Bartholdi, Craig A Tovey, and Michael A Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and welfare*, 6(2):157–165, 1989.

[4] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[5] Caitlin Kuhlman and Elke Rundensteiner. Rank aggregation algorithms for fair consensus. *Proceedings of the VLDB Endowment*, 13(12), 2020.

[6] Kathleen Cachel, Elke Rundensteiner, and Lane Harrison. Mani-rank: Multiple attribute and intersectional group fairness for consensus ranking. In *2022 IEEE 38th Intl. Conf. on Data Engineering (ICDE)*. IEEE, 2022.

[7] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics*, 19(12):2277–2286, 2013.

[8] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T Brown, and Alex Endert. Podium: Ranking data using mixed-initiative visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):288–297, 2017.

[9] Weichen Liu, Sijia Xiao, Jacob T Browne, Ming Yang, and Steven P Dow. Consensus: Supporting multi-criteria group decisions by visualizing points of disagreement. *ACM Transactions on Social Computing*, 1(1):1–26, 2018.

[10] Emily Hindalong, Jordon Johnson, Giuseppe Carenini, and Tamara Munzner. Towards rigorously designed preference visualizations for group decision making.

In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 181–190. IEEE, 2020.

[11] Emily Hindalong, Jordon Johnson, Giuseppe Carenini, and Tamara Munzner. Abstractions for visualizing preferences in group decisions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–44, 2022.

[12] Niels Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*, 36(1):15–30, 2018.

[13] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[14] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[15] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[16] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[17] Tiankai Xie, Yuxin Ma, Jian Kang, Hanghang Tong, and Ross Maciejewski. Fairrankvis: A visual analytics framework for exploring algorithmic fairness in graph mining models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):368–377, 2021.

[18] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. A nutritional label for rankings. In *Proceedings of the 2018 international conference on management of data*, pages 1773–1776, 2018.

[19] Yongsu Ahn and Yu-Ru Lin. Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics*, 26(1):1086–1095, 2019.

[20] Weina Jin, Diane Gromala, Carman Neustaedter, and Xin Tong. A collaborative visualization tool to support doctors' shared decision-making on antibiotic

prescription. In *Companion of the 2017 ACM Conf. on Computer Supported Cooperative Work and Social Comput.*, pages 211–214, 2017.

[21] Afra Mashhadi, Annuska Zolyomi, and Jay Quedado. A case study of integrating fairness visualization tools in machine learning education. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[22] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[23] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[24] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009.

[25] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[26] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439, 2022.

[27] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[28] S Bajracharya, Giuseppe Carenini, B Chamberlain, K Chen, D Klein, David Poole, Hamed Taheri, and Gunilla Öberg. Interactive visualization for group decision analysis. *International Journal of Information Technology & Decision Making*, 17(06):1839–1864, 2018.

[29] Giuseppe Carenini and John Loyd. Valuecharts: analyzing linear models expressing preferences and evaluations. In *Proceedings of the working conference on Advanced visual interfaces*, pages 150–157, 2004.

[30] Evanthia Dimara, Paola Valdivia, and Christoph Kinkeldey. Dcpairs: A pairs plot based decision support system. In *EuroVis-19th EG/VGTC Conference on Visualization*, 2017.

[31] Paul Hansen and Franz Ombler. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *Journal of Multi-Criteria Decision Analysis*, 15(3-4):87–107, 2008.

[32] Quantin Hayez, Yves De Smet, and Jimmy Bonney. D-sight: a new decision making software to address multi-criteria problems. *International Journal of Decision Support System Technology (IJDSST)*, 4(4):1–23, 2012.

[33] Sungsoo Hong, Minhyang Suh, Nathalie Henry Riche, Jooyoung Lee, Juho Kim, and Mark Zachry. Collaborative dynamic queries: Supporting distributed small group decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[34] Phi Giang Pham and Mao Lin Huang. Qstack: Multi-tag visual rankings. *J. Softw.*, 11(7):695–703, 2016.

[35] Di Weng, Ran Chen, Zikun Deng, Feiran Wu, Jingmin Chen, and Yingcai Wu. Srvis: Towards better spatial integration in ranking visualization. *IEEE transactions on visualization and computer graphics*, 25(1):459–469, 2018.

[36] Di Weng, Heming Zhu, Jie Bao, Yu Zheng, and Yingcai Wu. Homefinder revisited: Finding ideal homes with reachability-centric multi-criteria decision making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[37] Chirag Shah. Collaborative information seeking. *Journal of the Association for Information Science and Technology*, 65(2):215–236, 2014.

[38] Jyri Mustajoki and Raimo P Hämäläinen. Web-hipre: Global decision support by value tree and ahp analysis. *INFOR: Information Systems and Operational Research*, 38(3):208–220, 2000.

[39] Arthur H Copeland. A reasonable social welfare function. Technical report, Mimeo, University of Michigan USA, 1951.

[40] John G Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

[41] Markus Schulze. The schulze method of voting. *arXiv preprint arXiv:1804.02973*, 2018.

[42] Jean-Charles de Borda et al. Mathematical derivation of an election system. *Isis*, 44(1-2):42–51, 1781.

[43] Michael Behrisch, James Davey, Svenja Simon, Tobias Schreck, Daniel Keim, and Jörn Kohlhammer. Visual comparison of orderings and rankings. In *EuroVis*, 2013.

[44] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[45] Brittany Johnson and Yuriy Brun. Fairkit-learn: a fairness evaluation and comparison toolkit. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, pages 70–74, 2022.

[46] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56. IEEE, 2019.

[47] David Munechika, Zijie J Wang, Jack Reidy, Josh Rubin, Krishna Gade, Krishnaram Kenthapadi, and Duen Horng Chau. Visual auditor: Interactive visualization for detection and summarization of model biases. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 45–49. IEEE, 2022.

[48] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 473–484, 2022.

[49] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 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*, pages 99–106, 2019.

[50] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[51] Megha Srivastava, Hoda Heidari, and Andreas Krause. 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*, pages 2459–2468, 2019.

[52] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 392–402, 2020.

[53] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. Who gets what, according to whom? an analysis of fairness perceptions in service allocation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 555–565, 2021.

[54] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[55] Heidi Lam, Melanie Tory, and Tamara Munzner. Bridging from goals to tasks with design study analysis reports. *IEEE trans. on visualization and computer graphics*, 24(1):435–445, 2017.

[56] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013.

[57] Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone, Jim Davies, and Min Chen. Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2603–2612, 2012.

[58] Julian Heinrich and Daniel Weiskopf. State of the art of parallel coordinates. In *Eurographics (State of the Art Reports)*, pages 95–116, 2013.

[59] Michael Blumenschein, Xuan Zhang, David Pomerenke, Daniel A Keim, and Johannes Fuchs. Evaluating reordering strategies for cluster identification in parallel coordinates. In *Computer Graphics Forum*, volume 39, pages 537–549. Wiley Online Library, 2020.

[60] Carolina Nobre, Dylan Wootton, Lane Harrison, and Alexander Lex. Evaluating multivariate network visualization techniques using a validated design and crowdsourcing approach. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[61] Royce Kimmons. Exam scores, 2012.

[62] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36, 2019.

[63] Hilson Shrestha, Kathleen Cachel, Mallak Alkhathlan, Elke Rundensteiner, and Lane Harrison. Fairfuse: Interactive visual support for fair consensus ranking. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 65–69. IEEE, 2022.

[64] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[65] Rareş Constantin, Moritz Dück, Anton Alexandrov, Patrik Matošević, Daphna Keidar, and Mennatallah El-Assady. How do algorithmic fairness metrics align with human judgement? a mixed-initiative system for contextualized fairness assessment. In *2022 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pages 1–7. IEEE, 2022.