Help or Hinder? Evaluating the Impact of Fairness Metrics and Algorithms in Visualizations for Consensus Ranking

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

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

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

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

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

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 a study to analyze the effectiveness of integrating such fairness metrics-based visualization and algorithms. We explore this through a task-based crowdsourced experiment comparing an interactive visualization system for constructing consensus rankings, ConsensusFuse, with a similar system that includes visual encodings of fairness metrics and fair-rank generation algorithms, FairFuse. We analyze the measure 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 these 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 and along with emerging directions for future research.

CCS CONCEPTS

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

KEYWORDS

fairness, visualization, empirical study

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594108

ACM Reference Format:
Hilson Shrestha, Kathleen Cachel, Mallak Alkhathlan, Elke Rundensteiner, and Lane Harrison. 2023. Help or Hinder? Evaluating the Impact of Fairness Metrics and Algorithms in Visualizations for Consensus Ranking. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3593013.3594108

1 INTRODUCTION

The broader fairness community has developed a vast array of metrics and algorithms that conceptualize, measure, and systematize definitions of fairness, in part to guide decision-making in computing contexts. One popular medium for operationalizing these metrics in user-centric computing systems are interactive visualizations. Such visualizations can provide increased transparency across the underlying data, the decision algorithms applied to the data, and the corresponding fairness properties expressed by fairness metrics, among other benefits.

Several recent efforts highlight the inherent promise of interactive visualization for advancing goals in the fairness community, such as work from Mitchell *et al.* and Crisan *et al.* on model cards [15, 36], and Van Berkel *et al.* on examining the value of visualization over text for communicating fairness concepts [51]. However, efforts combining visualization approaches and fairness metrics and algorithms raise both challenges as well as unique opportunities in this space. 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 paper, we explore these broad questions through a particular instance of a controlled task-based visualization study. Our context is fair ranking problems, where prior works have focused on achieving group fairness, i.e. treating all groups in the ranking similarly, *e.g.* [1, 10, 30, 48, 57]. We begin with a recently published fairness visualization with an available open source system, FairFuse [48], which visualizes both the candidate items to be ranked and various fairness metrics such as Favored Pair Representation (FPR), Attribute Rank Parity (ARP), and PD Loss [10]. We adapt FairFuse into a new system, ConsensusFuse, by removing visualized

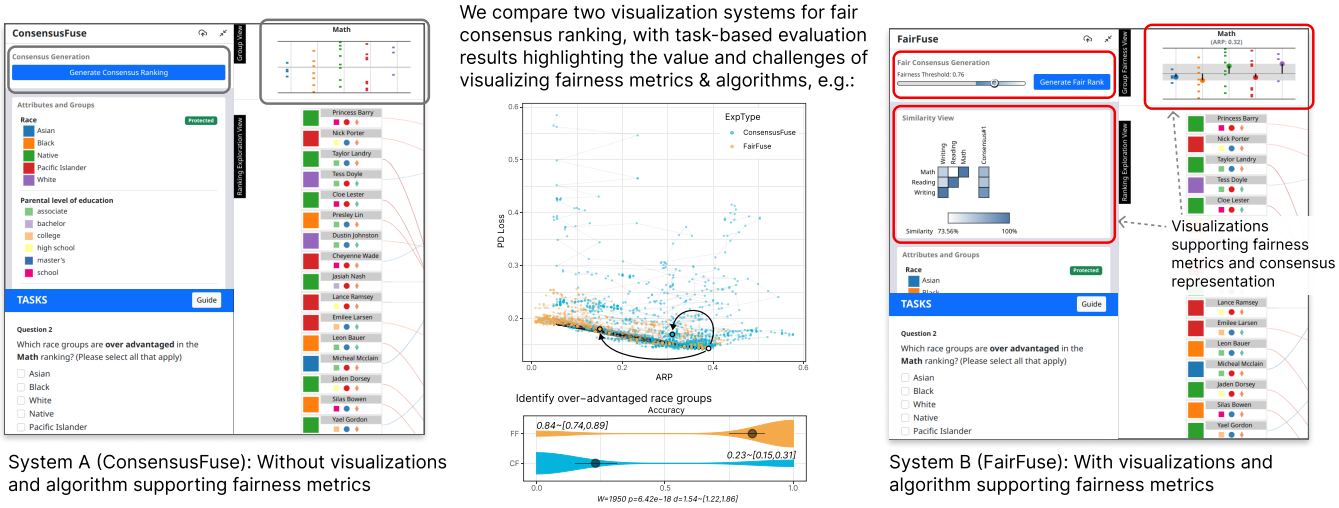


Figure 1: Can visualization-enabled fairness metrics aid fairness related tasks? We compare two systems: A: ConsensusFuse, a visualization that enables fairness comparison only by interactive visual displays of underlying items. B: FairFuse, a similar system which visualizes additional fairness metrics and provides a fair-rank generation algorithm. We find positive impact of embedding fairness metrics and algorithms into visualization supporting consensus ranking scenarios, but also certain risks.

fairness metrics and algorithms (Figure 2). We conduct a goal and activity analysis (*e.g.* [23, 24, 39] to define fairness-oriented tasks in ranking contexts (Table 1, 2). We distill these goals and activities into a set of evaluation tasks, with measurable outcomes (Table 3).

With the two systems and the above-identified tasks in place, we conduct a controlled experiment with $n = 200$ participants (100 per condition). Results generally validate that visualizing fairness metrics leads to notably increased accuracy in key fairness-related tasks. However, deeper analysis of measures, exploration behavior, and participant explanations reveal nuance, challenges and risks in visualizing fairness metrics. We discuss findings, such as for instance 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.

Contributions. Taken together, our work makes the following contributions:

- A task-based evaluation comparing a system that visualizes fairness metrics/algorithm results against a control with equivalent functionality, sans metrics/algorithms.
- Results that generally validate the value of visualizing fairness metrics/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.

2 RELATED WORK

Research to date has developed several fairness-related visualization systems, such as tools for consensus-building and ranking-based tasks. However, throughout these efforts, there remains a lack of empirical evidence examining the value and challenges of visualizing and incorporating fairness metrics and algorithms into human-in-the-loop systems. Here we review several of the systems, efforts, and concepts we draw from when designing the present experiment.

2.1 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.

2.1.1 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 [3, 6, 44, 56, 57]. 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 [5, 26, 50]. Crisan *et al.* and Mitchell *et al.* [15, 36] 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 [9, 38]. In the context of rankings, Yang *et al.* [58] provided "nutritional facts" for the fairness of rankings, Ahn *et al.* [1] proposed an interactive system for building fair rankings, and Xie *et al.* [57] introduced a visual system for fairness comparing rankings produced from graph mining recommender algorithms.

2.1.2 Evaluation of Fairness-Oriented Toolkits. Several researchers assessed toolkits that incorporate fairness into their process. Mashhadi *et al.* [35] studied the impact of the visualization styles of six open-source fair classification toolkits on student learning of fairness criteria. Lee *et al.* [32] 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 [56], which were designed for broader audiences. Richardson *et al.* [43] 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.* [16] have highlighted gaps between fairness toolkits' capabilities and practitioners' needs.

2.1.3 Evaluation on Presentation of Fairness Information. Studies have evaluated the presentation of fairness related information in different scenarios. Van Berkel *et al.* [51] 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.* [45] 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.* [12] 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.* [49] found that people prefer demographic parity among the 6 different notions of group fairness. Harrison *et al.* [21] 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.* [19] 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.

2.2 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 [2, 11, 17, 20, 22, 25, 33, 40, 42, 47, 54, 55]. A subset of these tools consider the setting, like ours, in which stakeholder preferences are encoded as rankings [11, 23, 33]. Liu *et al.* [33] 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.* [24] 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 [11, 23, 33]. The corresponding evaluation studies are focused on how well these tools help achieve consensus outcomes – yet none consider the employment of consensus generation algorithms [7, 14, 27, 46]. 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.

2.3 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.* [18] 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.* [53] 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.* [4] 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.

3 VISUALIZATION AND INTERACTION DESIGN

To be able to study the challenges and opportunities in visualizing fairness metrics and algorithms, we have modified the FairFuse system [48] to create two system variations. The FairFuse system employed task abstraction methodologies following procedures from Lam *et al.* [31] and recent works on group decision-making by Hindalong *et al.* [23, 24]. Table 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 2) based on a widely used method in the visualization literature [8], leading to the design and implementation of several views.

3.1 FairFuse

The FairFuse system (Figure 1B) consists of several views to support the goals (Table 1) and activities (Table 2).

- **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 2F. Each candidate's set of attributes and values are represented by glyphs and colors [34] (A5, A6), collectively called a Candidate Card. By dragging-and-dropping the Candidate Card (A18), any generated consensus ranking can be adjusted if necessary.
- **Group Fairness View** (Figure 2E) captures fairness of a ranking at individual group level utilizing FPR score [10] (A6, A7, A8, A9, A10) and holistically across groups in the ranking using ARP score [10] (A11, A12, A14).

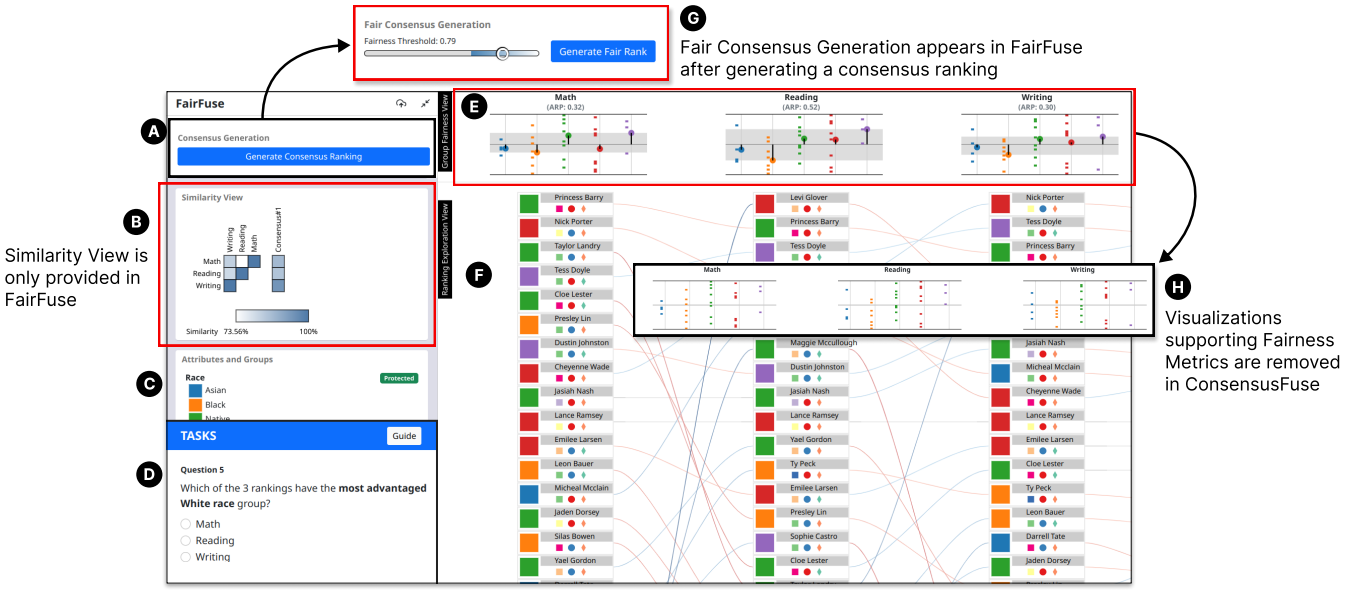


Figure 2: FairFuse and ConsensusFuse System Designs with changes in visualizations related to fair consensus generation. A) Consensus Generation, B) Similarity View (in FairFuse), C) Attributes Legend, D) Tasks presented to the participants, E) Group Fairness View (in FairFuse), F) Ranking Exploration View, G) Fair Consensus Generation (in FairFuse), H) Group View (in ConsensusFuse).

Table 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

- **Similarity View** (Figure 2B) 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 [28].
- **Ranking Generation** process uses a button to first generate a consensus ranking without any fairness intervention 2A. After the consensus ranking is displayed, the generation button is replaced with a slider (Figure 2E) – allowing the fairness threshold of generated consensus ranking to be adjusted (A13, A18). This process utilizes the Fair-Copeland algorithm [10].

3.2 ConsensusFuse

An alternate and functionally equally capable version of FairFuse was created, called ConsensusFuse(Figure 1A), which 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 2H), 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 [10]. Differences are shown in Figure 2.

4 STUDY DESIGN

We aim to investigate the challenges and opportunities of a system like FairFuse for the activities associated with fairness-oriented tasks. 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

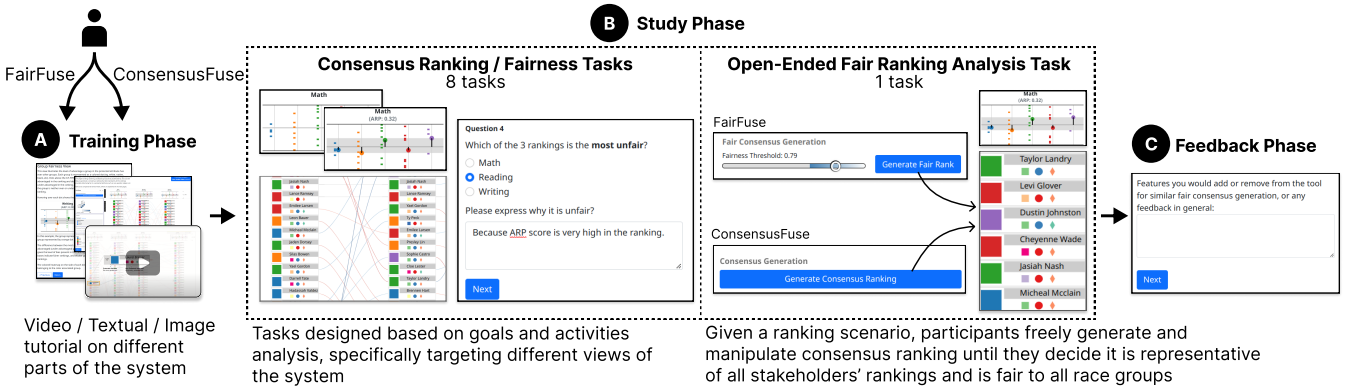


Figure 3: Study Design: We explore using visualization-enabled fairness metrics in building a fair consensus ranking. Participants are divided into two conditions, FairFuse: system with visualization-enabled fairness metrics and ConsensusFuse: system without visualizations for fairness metrics. Participants go through three phases. A) Training Phase, B) Study Phase and C) General Feedback Phase.

a between-subjects study in which each participant was assigned to use either FairFuse or ConsensusFuse system.

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 3) as seen in a similar user study in the literature [41].

Training Phase. The study starts with the training phase (Figure 3A) 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.

Study Phase. The second phase (Figure 3B) 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 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 3C) 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 [29] 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 2), we created a set of tasks for the participants, listed in Table 3. 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.

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. Random assignment was achieved through round-robin online recruitment using the Prolific platform. Prolific reporting shows that 99 participants (separate from the 200 completions) returned the experiment before completing it. (On Prolific, participants can discontinue the experiment for any reason.) Beyond these, 6 participants in total timed out. 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).

Table 2: Activities resulting from the goals and activities analysis, designed to support the goals in Table 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 [28] 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

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 4. 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 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 5a) with a large effect size ($d = 1.54 \sim [1.22, 1.86]$). The careful design of the Group 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 5b.

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 6a and 6b). 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 [10], **T4** focused on utilizing the ARP metric [10]. 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 5c. We also instructed participants to express why they think their ranking choice is unfair. Participant comments reflect at least two types of reasoning 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.*

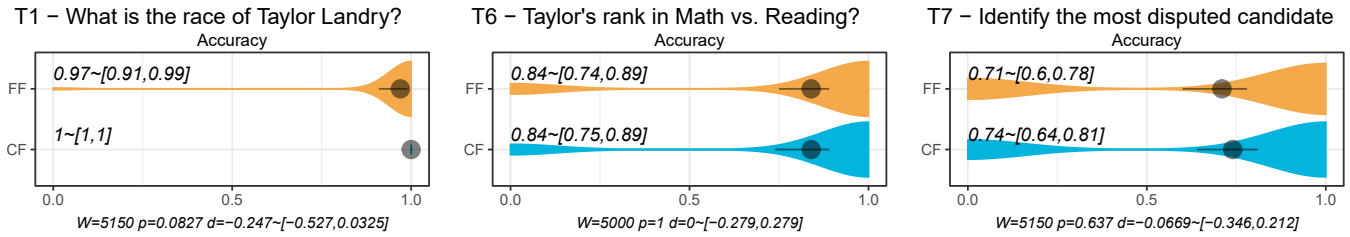
Expression at the understanding level goes beyond just reporting the ARP score, such as:

> *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.*

In contrast, some ConsensusFuse participants considered only a single group resulting in incorrect answers, such as:

Table 3: List of task prompts given to the participants. Tasks are targeted at the *Goals and Activities* (Table 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



(a) Locating protected attribute [T1] **(b) Utilizing PCP Position Comparison [T6]** **(c) Interpreting PCP Gradient [T7]**

Figure 4: Results for Ranking Exploration tasks

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

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

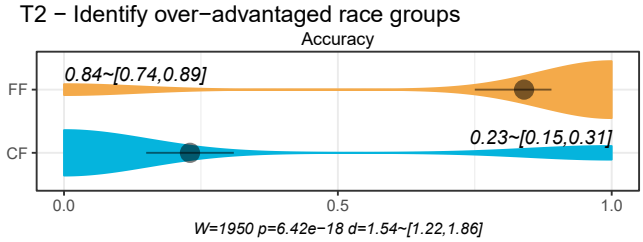
> *100% of the white students are in the top half, but only 28.5% of the black students are.*

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 5d. 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 6b).

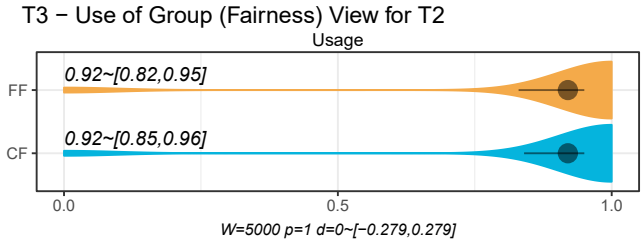
5.3 Consensus Representation and Analysis Tasks

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 7) 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.

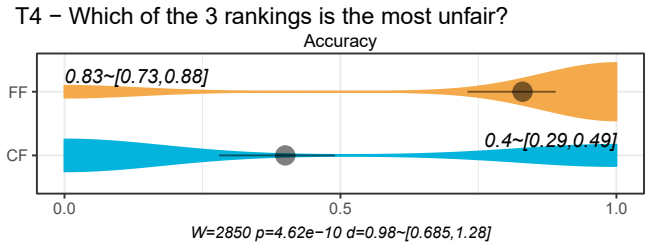
For this particular task, we also asked participants elaborate on their answers. We find that some participants in FairFuse, even



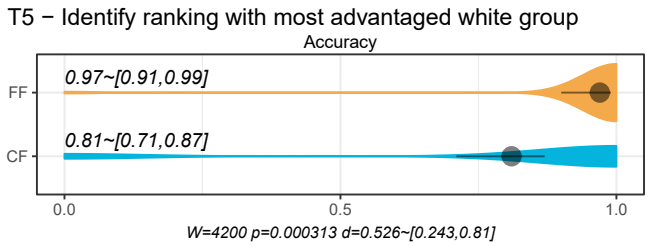
(a) Results for Identifying advantaged Group(s) [T2] (both native and white as correct answer)



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



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

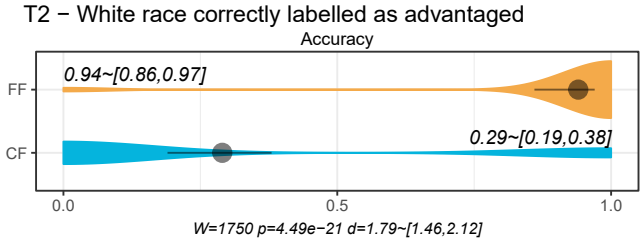


(d) Results for Identifying Group-level Unfairness [T5]

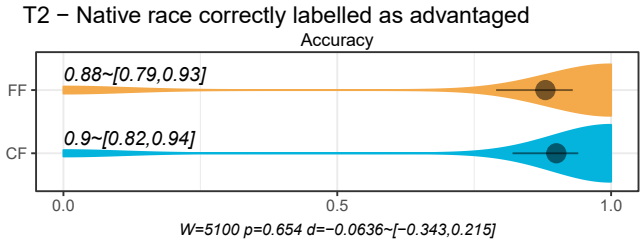
Figure 5: Results for Fairness-oriented tasks

though they correctly identify the most dissimilar ranking, mention using the Group Fairness View instead of the Similarity View, such as:

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



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

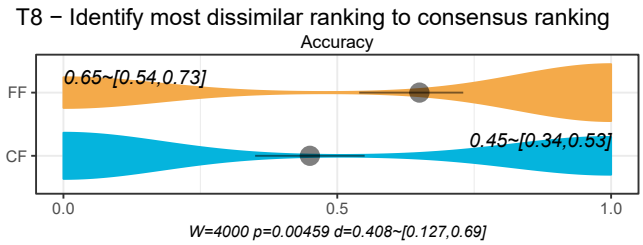


Figure 7: Results on Using Consensus Generation Procedure [T8]

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 quantitative data in aggregate supports the notion that FairFuse performs better in identifying the (dis)similarity between the consensus ranking and base rankings, qualitative results do not fully support this conclusion. Participants may focus more on the fairness metrics compared to other available 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. We analyze the ARP scores between the two groups (which ranges from 0 to 1, with 0 representing a ranking with perfect statistical parity [10]) 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 [10] (representation of base rankings in the consensus ranking, with 0 representing that all the base rankings exactly match the consensus ranking) in both groups are similar (Figure 8b) 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 9D). Figure 9A 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 9B marks the vastly improved mean ARP value in FairFuse compared to Figure 9C in ConsensusFuse. We find that FairFuse participants make fewer interactions to agree on a fair consensus ranking as shown in the violin plot (Figure 8c) (ConsensusFuse: $M = 18.76 \sim [15.03, 24.62]$ vs. FairFuse: $M = 12.2 \sim [10.1, 15.08]$).

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.

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

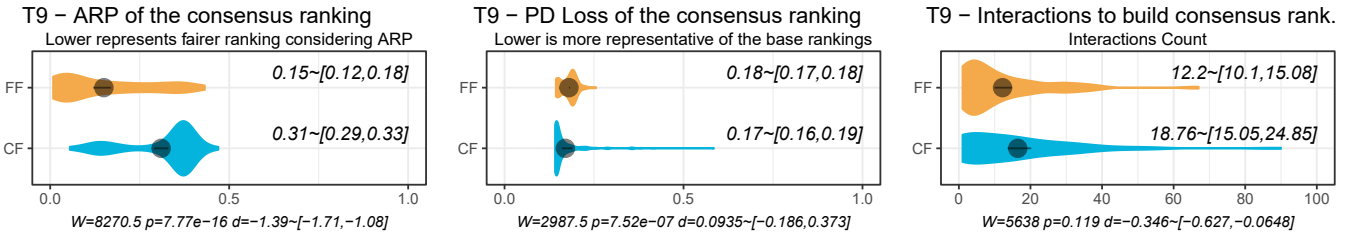
Given evolving societal norms and values, definitions of fairness can change over time and place. Definitions also vary from one discipline to another [37]. Algorithms designed to assist in incorporating fairness incorporate ways for decision-makers to tune fairness in the specific problem context. This 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.

6.2 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 tend to make decisions that are consistent with visualization-enabled fairness metrics Figure 9. 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.

6.3 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 6). 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.



(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 8: Results for Open-Ended Fair Ranking Analysis Task

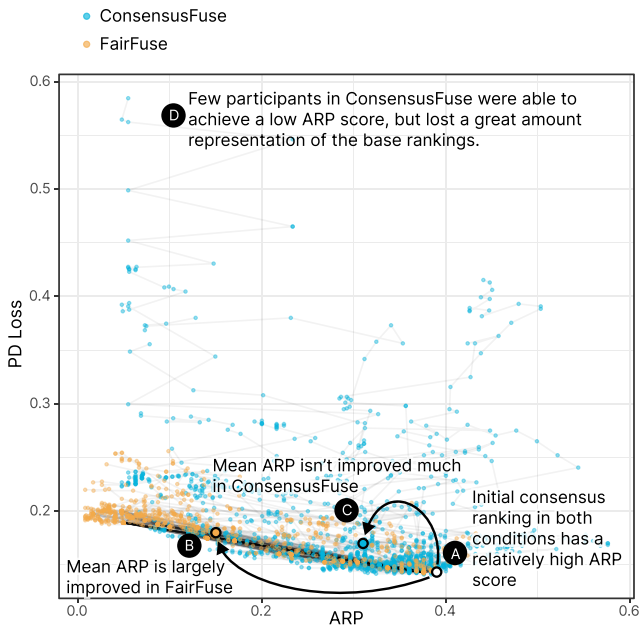


Figure 9: Results of ARP vs. PD Loss throughout each user interaction while generating a fair consensus ranking. The white dot indicates the ARP and PD Loss of the initial consensus ranking in both conditions.

6.4 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.

7 LIMITATIONS AND FUTURE WORK

The limitations of FairFuse [48] are also relevant to this crowdsourced study. FairFuse focusses on ARP and FPR fairness metrics [10] 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 [52] 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 [13]. Therefore, future work could incorporate multiple fairness definitions 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.

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 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.

ACKNOWLEDGMENTS

This work was supported in part by NSF IIS #2007932.

REFERENCES

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

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

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

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

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

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

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

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

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

[10] Kathleen Cachel, Elke Rundensteiner, and Lane Harrison. 2022. MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking. In *2022 IEEE 38th Intl. Conf. on Data Engineering (ICDE)*. IEEE.

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

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

[13] Rareş Constantin, Moritz Dück, Anton Alexandrov, Patrik Matošević, Daphna Keidar, and Mennatallah El-Assady. 2022. 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)*. IEEE, 1–7.

[14] Arthur H Copeland. 1951. *A reasonable social welfare function.* Technical Report. Mimeo, University of Michigan USA.

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

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

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

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

[19] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. 2021. 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*. 555–565.

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

[21] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 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*. 392–402.

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

[23] Emily Hindalong, Jordon Johnson, Giuseppe Carenini, and Tamara Munzner. 2020. Towards Rigorously Designed Preference Visualizations for Group Decision Making. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 181–190.

[24] Emily Hindalong, Jordon Johnson, Giuseppe Carenini, and Tamara Munzner. 2022. Abstractions for Visualizing Preferences in Group Decisions. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–44.

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

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

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

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

[29] Royce Kimmons. 2012. Exam scores. http://roycekimmons.com/tools/generated_data/exams

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

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

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

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

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

[35] Afra Mashhadi, Annuska Zolyomi, and Jay Quedado. 2022. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

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

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

[38] David Munechika, Zijie J Wang, Jack Reidy, Josh Rubin, Krishna Gade, Krishnaram Kenthapadi, and Duen Horng Chau. 2022. Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases. In *2022 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 45–49.

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

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

[41] Carolina Nobre, Dylan Wootton, Lane Harrison, and Alexander Lex. 2020. 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*. 1–12.

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

[43] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. 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*. 1–13.

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

[45] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[46] Markus Schulze. 2018. The Schulze method of voting. *arXiv preprint arXiv:1804.02973* (2018).

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

[48] Hilson Shrestha, Kathleen Cachel, Mallak Alkhathlan, Elke Rundensteiner, and Lane Harrison. 2022. FairFuse: Interactive Visual Support for Fair Consensus Ranking. In *2022 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 65–69.

[49] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2459–2468.

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

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

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

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

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

[55] Di Weng, Heming Zhu, Jie Bao, Yu Zheng, and Yingcai Wu. 2018. 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*. 1–12.

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

[57] Tiankai Xie, Yuxin Ma, Jian Kang, Hanghang Tong, and Ross Maciejewski. 2021. FairRankVis: A Visual Analytics Framework for Exploring Algorithmic Fairness in Graph Mining Models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 368–377.

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

A DEMOGRAPHICS AND RESULTS SUMMARY

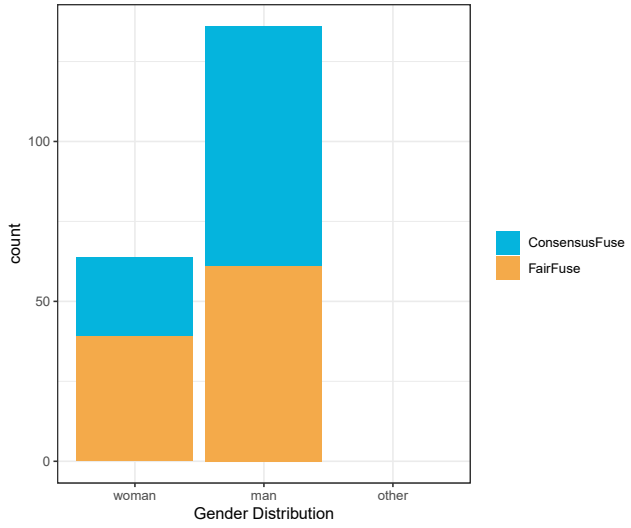


Figure 10: Gender Distribution

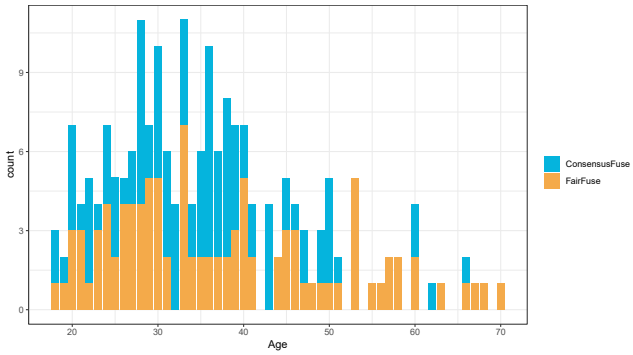


Figure 11: Age Distribution

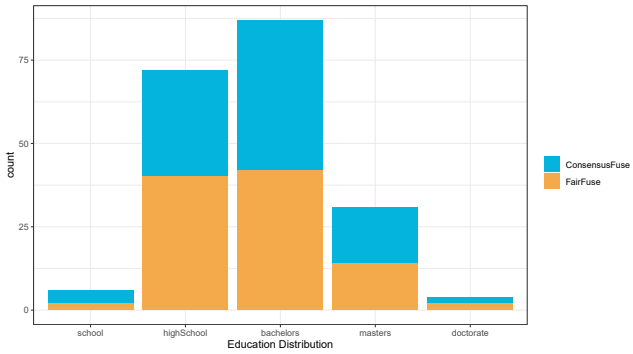


Figure 12: Education Level Distribution

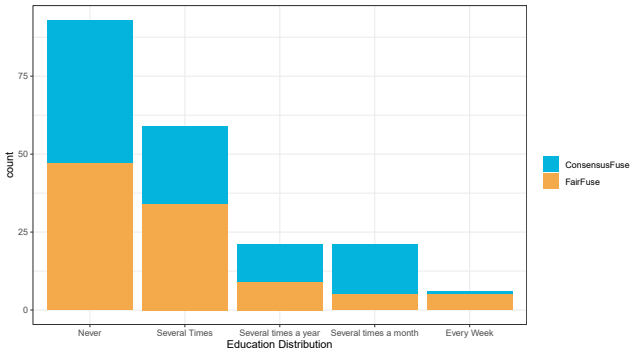


Figure 13: Visualization Experience Distribution

Table 4: Summary of the results by tasks

	Task	FairFuse	ConsensusFuse	Result
T1	Locating protected attribute	$0.97 \sim [0.91, 0.99]$	$1 \sim [1, 1]$	$W = 5150\ p = 0.0827\ d = -0.247 \sim [-0.527, 0.0325]$
T2	Identifying Advantaged Group(s)	$0.84 \sim [0.74, 0.89]$	$0.23 \sim [0.15, 0.31]$	$W = 1950\ p = 6.42e-18\ d = 1.54 \sim [1.22, 1.86]$
	a) Correctly identify race 1	$0.94 \sim [0.86, 0.97]$	$0.29 \sim [0.19, 0.38]$	$W = 1750\ p = 4.49e-21\ d = 1.79 \sim [1.46, 2.12]$
	b) Correctly identify race 2	$0.88 \sim [0.79, 0.93]$	$0.9 \sim [0.82, 0.94]$	$W = 5100\ p = 0.654\ d = -0.0636 \sim [-0.343, 0.215]$
T3	Visualization Use	$0.92 \sim [0.82, 0.95]$	$0.92 \sim [0.85, 0.96]$	$W = 5000\ p = 1\ d = 0 \sim [-0.279, 0.279]$
T4	Identifying Attribute-level Unfairness	$0.83 \sim [0.73, 0.88]$	$0.4 \sim [0.29, 0.49]$	$W = 2850\ p = 4.62e-10\ d = 0.98 \sim [0.685, 1.28$
T5	Identifying Group-level Unfairness	$0.97 \sim [0.91, 0.99]$	$0.81 \sim [0.71, 0.87]$	$W = 4200\ p = 0.000313\ d = 0.526 \sim [0.243, 0.81]$
T6	Utilizing PCP Position Comparison	$0.84 \sim [0.74, 0.89]$	$0.84\ [0.75, 0.89]$	$W = 5000\ p = 1\ d = 0 \sim [-0.279, 0.279]$
T7	Interpreting PCP Gradient	$0.71 \sim [0.6, 0.78]$	$0.74 \sim [0.64, 0.81]$	$W = 5150\ p = 0.637\ d = -0.0669 \sim [-0.346, 0.212]$
T8	Using Consensus Generation Procedure	$0.65 \sim [0.54, 0.73]$	$0.45 \sim [0.34, 0.53]$	$W = 4000\ p = 0.00459\ d = 0.408 \sim [0.127, 0.69]$
T9	Using Fair Consensus Generation Procedure			
	a) ARP	$0.15 \sim [0.12, 0.18]$	$0.31 \sim [0.29, 0.33]$	$W = 8270.5\ p = 7.77e-16\ d = -1.39 \sim [-1.71, -1.08]$
	b) PD Loss	$0.18 \sim [0.17, 0.18]$	$0.17 \sim [0.16, 0.19]$	$W = 2987.5\ p = 7.52e-07\ d = 0.0935 \sim [-0.186, 0.373]$
	c) Interactions Count	$12.2 \sim [10.1, 15.08]$	$18.76 \sim [15.05, 24.85]$	$W = 5638\ p = 0.119\ d = -0.346 \sim [-0.627, -0.0648]$