rutaBAGA: A Visualization Approach for Bias Awareness in University Admissions

YANAN DA, YUTONG BU, YILING LI, and EMILY WALL, Emory University, USA

Introduction: University admissions is a complex decision making process where implicit biases may impact the way reviewers individually and collectively make decisions. Education-based methods like training courses often have limited impact due to the subconscious nature of these biases. This paper introduces a visualization system, rutaBAGA, that promotes heightened awareness of implicit biases through real-time system interactions.

Data collection: Using our system, we conducted (i) a pre-registered crowdsourced user study where 75 participants performed a simulated review of undergraduate university applications and (ii) a case study with the Ph.D. admissions committee in the Computer Science department at a private university. We collected participants' interaction logs with the system (to derive review time and behavior/decision changes), ratings on the applications, and post-study survey responses in both studies. In addition, we collected Implicit Association Test (IAT) scores from the crowdsourced study and interview recordings from the case study.

Data Analysis: For the crowdsourced study, we conducted (i) t-tests to compare participants' time spent and competitiveness ratings on applicants across different gender/racial groups and (ii) linear regressions to analyze the relationship between implicit gender/racial bias (as measured by IAT scores) and differences in review behaviors (time spent) and decisions (competitiveness ratings). For both of the studies, we analyzed the interaction logs and survey responses/interview transcripts to understand the system's impact on participants review process.

Results: The results of the crowdsourced study showed that implicit racial bias correlates to differences in review behaviors and decisions. The results from both the crowdsourced study and the case study showed that our system can increase awareness of undesired behaviors and lead to behavioral/decision changes for some participants.

Materials: The source code for the system and the user study materials can be found at https://osf.io/t6hjd/.

Conclusion: We presented a visualization system that enables reviewers to scrutinize their own processes to ensure fair and consistent review procedures. Results from a crowdsourced study showed (i) implicit racial bias correlates to observable differences in university application review behaviors and decisions and (ii) our system can affect individuals' review processes. Results of a case study with the Computer Science department at a private university demonstrated rutaBAGA can facilitate bias-aware decision making in a real-world system deployment.

CCS Concepts: • **Human-centered computing** → **Visual analytics**.

Additional Key Words and Phrases: bias awareness, implicit biases, university admissions, decision making

ACM Reference Format:

Yanan Da, Yutong Bu, Yiling Li, and Emily Wall. 2018. rutaBAGA: A Visualization Approach for Bias Awareness in University Admissions. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 24 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

1 INTRODUCTION

University admissions involves complex decision making processes, often characterized by individual reviewers reading application packets and rating them based on factors such as the applicant's academic performance, non-academic accomplishments, and personal qualities such as communication skills. Reviewers' individual evaluations are often then collated and discussed among a committee to inform final admissions decisions. Given the complexity and subjectivity of this process, **unconscious biases** (biases beneath the level of an individual's conscious attention) might influence reviewers' decisions.

This work focuses on a specific type of unconscious bias – implicit biases, which can include gender and racial bias, that are ingrained as a result of cultural norms or past experiences [21, 22]. Unlike conscious biases such as overt sexism or racism which involve deliberate prejudice against people of a certain group, implicit biases are not intentional and may not always align with one's explicit beliefs [45]. Nevertheless, they often materialize in harmful ways such as snap judgments based on an individual's skin color or gender. Importantly, implicit biases are not conscious, they often persist despite an individual's best efforts to counteract it. Our goal in this work is therefore to investigate (i) to what extent implicit biases correspond to observable behaviors in the review process, (ii) how we can promote real-time reflective review processes, and (iii) to what extent heightened awareness leads to reviewers' adjustment of associated behaviors and decisions.

We designed an interactive system, **RUTABAGA**, to support a **b**ias **a**ware **g**raduate **a**dmissions process. We introduced the intervention in the context of graduate admissions and later generalized the concept to undergraduate admissions as well but retained the system name. In this paper, we define **bias** to encompass inequities in (i) the review process (via disparities in time spent reviewing applicants of different racial and gender groups) and (ii) decisions (via disparities in perceived competitiveness of applicants of different racial and gender groups). We aim to achieve **awareness** by surfacing these disparities to reviewers through visualizations to encourage real-time **self-reflection** on the review process.

Our primary contributions include:

(1) the system, RUTABAGA, designed in collaboration with two Ph.D. program admissions committee chairs to promote reflection on review processes and increase reviewer awareness of undesired behaviors,
(2) empirical results demonstrating (i) the relationship between implicit biases and observable review behaviors and (ii) the impact of RUTABAGAon bias awareness,
(3) results of a case study with the Computer Science department at a private university that demonstrate how RUTABAGA can facilitate bias-aware decision making in a real-world system deployment.

We discuss the implications of our findings, and in particular, how they relate to the recent Supreme Court decision on affirmative action that ruled universities may no longer make admissions decisions on the basis of race [38]. We conclude with a call-to-action – that research to address biases in admissions processes has never been more urgent.

2 RELATED WORK

We discuss how increasing bias awareness during university admissions can empower reviewers to address potential **biases in admissions** (Section 2.1), informed by recent work on **biases in visualization** (Section 2.2).

2.1 Biases in Admissions

Research in higher education has examined the forms of bias that can impact decision making processes including university admissions [40]. These biases influence decisions in ways that can systematically disadvantage specific groups. Woo et al. [58] evaluated the validity, bias, and fairness of the main source of information used in graduate admission decision making, such as GPA, GRE, personal statement, and identified that each of the assessments is associated with potential biases. For example, qualitative assessments like personal statement are subject to sociocognitive and rater biases [58].

Implicit biases [22] which are unconscious preferences reflecting implicit attitudes or stereotypes have received considerable attention due to their prevalence and invisibility [40]. For example, one laboratory experiment showed that science faculty exhibit a bias against female students such that male candidates are rated as significantly more competent and hireable than female candidates with otherwise identical credentials [34]. Experimental results show similar racial discrimination in the labor market: equivalent resumes with White names received 50% more callbacks for interviews than resumes with African-American names [4]. Pieper and Krsmanovic [39] interviewed graduate faculty members who serve on admissions committees to examine the presence of implicit bias in the admissions process. The interview revealed that faculty members have varying levels of awareness regarding the presence of implicit bias in the admissions process and recognize the importance of the admission committee to reduce implicit bias.

The implicit Association Test (IAT) [23] characterizes implicit biases by measuring the association that people hold between attributes and concepts. The test asks users to quickly and accurately categorize words or images and in turn measures reaction time, such that faster (correct) responses indicate stronger associations than slower responses, suggesting how implicit attitudes can influence cognitive processes and behaviors. Using the IAT, Capers et al. [9] showed the presence of significant white preference in medical school admissions. Diversity training [5] has been used to address implicit biases in organizational and educational settings (e.g., to improve attitudes toward women in STEM [28]). However, since these biases are unconscious, informing individuals about the existence of implicit biases has apparently limited impact during the decision making process [16].

Our work contributes to this ongoing body of literature by introducing a real-time intervention designed to mitigate reviewers' potential bias during the admissions process. Unlike traditional methods such as training or post-review evaluations, our system encourages immediate self-reflection through interactive visualizations.

2.2 Bias in Visualization

Bias has been actively studied in the visualization community recently, formalizing the types of bias relevant to visualization and visual analytics [15, 52]. Some efforts have examined the presence of particular types of bias in decision making processes with visualizations such as the attraction effect [14], priming and anchoring bias [11, 49, 50, 57], confirmation bias [35], and Dunning-Kruger Effect [10]. Other recent work proposed computational metrics that can be applied to user interactions with data to quantify bias in real-time [17, 20, 51]. Prior metrics (e.g., [51]) can be noisy and wrongfully signal false positive biases [24], however. Hence, for the context of the application review process, we choose a simplified approach based on surfacing measures of *focus*: the cumulative duration of time spent interacting with different applications and different application components, without assigning a value judgment to the outcomes.

Apart from bias detection, researchers have also recently investigated methods to mitigate bias [6, 29, 54] by altering the framing of the task [13] or communicating bias metrics visually in real-time to increase the awareness of bias [36, 53]. The design of our system is inspired by recent work [36, 53] that captures and visualizes users' interaction history with

data in real-time to promote reflection of one's data analysis process. While these efforts had mixed quantitative results on biased *decisions* in laboratory experiments, they did positively impact *awareness*.

Within the education space, visualization researchers have studied the holistic application review process and suggested possible designs for supporting complex decision making [31, 46, 47], such as presenting alternative visual representations of application attributes, integrating sensemaking and storytelling tools [32], and providing rating recommendations using machine learning methods [31]. Unique from these efforts, we explore a real-world deployment of these techniques and focus on implicit biases rather than cognitive biases.

Visualizations have also been employed for mitigating implicit bias in teaching and supporting equitable teaching practices [41–43]. The EQUIP tool [42] tracks students' verbal participation in the classroom to provide visualizations about the participation patterns across different demographic groups, allowing instructors to identify potential bias and make changes in their practices accordingly [43]. Our system is similar to the EQUIP tool in terms of presenting patterns across different gender and racial groups.

Prior studies suggest that directly informing individuals of their biases often has limited effectiveness [16]. Instead, we hypothesize that a *real-time* intervention is a promising solution by enhancing reviewers' awareness of their decision-making processes and encouraging reflection. Visualization, in particular, leverages visual perception to help users identify patterns, making it a powerful tool for promoting awareness and mitigating bias in decision-making. Our proposed system thus emphasizes self-reflection through interactive visual analytics, aligned with prior work [36, 53], to drive more meaningful behavioral changes.

3 FORMATIVE DESIGN

The design and development of RUTABAGA followed a user-centered approach [2] that involved close collaboration with two Ph.D.admissions committee chairs in the Computer Science Department of a private university.

3.1 Methodology

We first conducted **semi-structured interviews** with two admissions committee chairs, C1 and C2, in the Computer Science department to understand their Ph.D. admissions process in detail. The 30-minute interview covered topics including the application format, data access, decision making criteria, and collaborative mechanisms. The interviews illuminated **characteristics of the existing review process** (Section 3.2) and the program's need to conduct meta-analysis on the admissions review process.

Based on our understanding of the program needs, we next sketched several solutions and built a **preliminary prototype** to ground the discussion in a second interview. More details of the preliminary interface are attached in Supplemental Materials. In the second interview, we provided a **demonstration of the interface design** to the chairs, followed by a semi-structured interview to gauge initial reactions to the design. The session lasted approximately 45 minutes. An additional survey followed afterward, including questions about the usefulness of system components, willingness to use such a system in the future, and an opportunity to suggest alternatives in free-form text. The main feedback we received included the chairs' desires to (1) visually incorporate interaction *time* on an application to allow assessment of time spent across applicants (the preliminary system included visual encoding of discrete interaction *count*) and (2) make the interface simpler to use by separating the review phase and self-reflection phase into separate tabs which were initially adjacent views in a singular screen in the prototype.

Additional interviews were conducted with one committee chair (C1) after two additional iterations of design to collect **ongoing feedback on subsequent iterations** of the system. Each session lasted about 30 minutes. The

system was updated according to feedback including (i) adding visual encodings for applicants' gender and race and (ii) visualizing reviewers' time spent on each *component* of an application, as described in Section 4. From the interviews and questionnaire responses, we summarized the needs for analyzing the admissions review process as **process awareness needs** (Section 3.3) and refined our **design goals** (Section 3.4) for the final system design.

Following the initial deployment of rutaBAGA, the system was further refined based on observations of its real-world usage. These ongoing refinements are discussed in Section 9.

3.2 Prior Review Process

The interviews illuminated the prior review process which can be summarised as follows. Applicant portfolios, in the form of single PDF file (one per applicant) were distributed to a minimum of two reviewers assigned by the committee chairs based on applicant research interests. Each reviewer rated the assigned applicants according to pre-defined criteria such as Research Preparedness and Communication Proficiency on a 0-5 scale and shared optional free-form feedback in a shared spreadsheet. After the initial review, applicants who were "above the bar" were interviewed via Zoom by at least one faculty member. The committee then met to discuss and make admissions decisions, referencing a sorted version of the spreadsheet of reviewer scores to anchor discussions.

3.3 Process Awareness Needs

Based on the interviews, we identified reviewers' needs for investigating (i) internal consistency in *time spent* across applicants and application components, and (ii) internal consistency in *ratings and decisions* across applicants, stratified by attributes such as race and gender. Inconsistency in time spent (i), for instance, could be not spending enough time on a certain application, neglecting a certain application component (e.g., not reading recommendation letters), or systematically spending more/less time on a certain group of applications across gender or race. While time spent alone is a noisy metric influenced by many factors, as suggested by the admission chairs, it would be a useful metric that can nonetheless provide reviewers with some point of reference to spark reflection.

The system should also encourage reviewers to reflect on the decision outcomes and check internal consistency in their ratings (ii). Inconsistency in ratings could be that applications with the same ratings in evaluation criteria received different overall ratings (e.g., competitive v. not competitive), or that a certain group of applications was systematically rated as more/less competitive across gender or race, etc.

3.4 Design Goals

Based on our formative design activities, we derived the following **design goals (DG)** to support the program's desired framework for admissions review needs for increasing process awareness.

DG1. Facilitate independent review of applications. The system should enable reviewers to *independently* rate applications across pre-defined dimensions.

DG2. Support assessment of independent review behavior. The system should enable individual reviewers to analyze their decision making processes and outcomes to increase awareness of undesired behaviors, such as inconsistency in time spent across applications.

DG3. Minimize the barrier to entry. The system should be visually simple and intuitive to increase adoption of the system over the status quo method for reviewing, and to ensure the system is usable by faculty members beyond the visualization domain.

The system was also designed with goals in mind to support collective *group* awareness and decision making; however, we defer discussion of these features to supplemental materials as they were not the focal point of our subsequent experiments.

4 SYSTEM

Based on our design goals, we developed a system RUTABAGA (Figure 1) mainly consisting of two separate tabbed pages, including a Rating Page where reviewers read and rate applications independently (DG1); and Summary Page where reviewers can see a summary of their process as shown from their interactions with applicants (DG2).

4.1 Rating Page

The Rating Page (Figure 1, top) is designed to support seamless completion of existing tasks involved in independent review of applications (DG1). It consists of the following components.

(A) Document Viewer shows PDF documents including personal statement, resume, letters of recommendation, transcript, and so on, organized into tabs. **(B) Profile View** shows tabular attributes of the applicant such as GPA, degree, major, etc. A set of default attributes are shown initially, and users can select/deselect attributes to be shown from a drop-down list if they would like to make their decision process blind to some sensitive attributes. **(C) Comments View** allows reviewers to leave comments about the respective application. **(D) Ratings View** allows reviewers to (i) rate the applicant on a set of factors (which can be pre-defined by the committee based on their review criteria) on a 0-5 scale and (ii) rate the overall competitiveness of the applicant on a 1-4 scale (1 = Not Competitive, 4 = Very Highly Competitive). Composing the application documents, comments, and ratings in a single interface makes it simpler to evaluate applications and adopt the system (DG3).

Interaction Logs record reviewers' time-stamped interactions including clicking and scrolling on each applicant and each component of the application (DG2). Page visibility changes (i.e., navigating to a different application or browser tab) are recorded in order to facilitate filtering out time periods that users are not focused on the interface. Furthermore, the system applies thresholds to filter out outlier time periods (that are too short or too long) in order to reduce noise in derived time spent. Inspired by prior work in analytic provenance [37], which demonstrates that user interaction is a powerful cue to learn about users, we used the interaction logs to derive visualizations of reviewers' review process to help them maintain self-awareness of their decision making process.

4.2 Summary Page

While the Rating Page serves as an interface for completing existing reviewing tasks, the Summary Page (Figure 1, middle) aims to enhance awareness of individual reviewers' processes (DG2). Reviewers can access this page at any time during the admissions review cycle. It maintains the **Profile, Comments** and **Ratings** Views (Figure 1 B, C, and D, respectively) from the Rating Page, but replaces the **Document** (A) panel with a data visualization panel (A.1), as shown in the middle of Figure 1.

(E) Filters provide controls for filtering data by numerical or categorical attributes. Beyond the Profile attributes (gender, test scores, etc.), users can also filter by their assigned ratings (for teaching preparedness, communication, etc.) and overall recommendation of applicants (e.g., to view only candidates they rated as Competitive).

(F) Interactive Scatterplot visualizes reviewed applications, with the ability to assign x- and y-axes from a drop-down list to represent variables like GRE score, GPA, reviewer's ratings, and overall recommendation. *Hovering* on a point (applicant) in the scatterplot populates the Profile, Comment, and Ratings Views, and the Time Spent Distribution

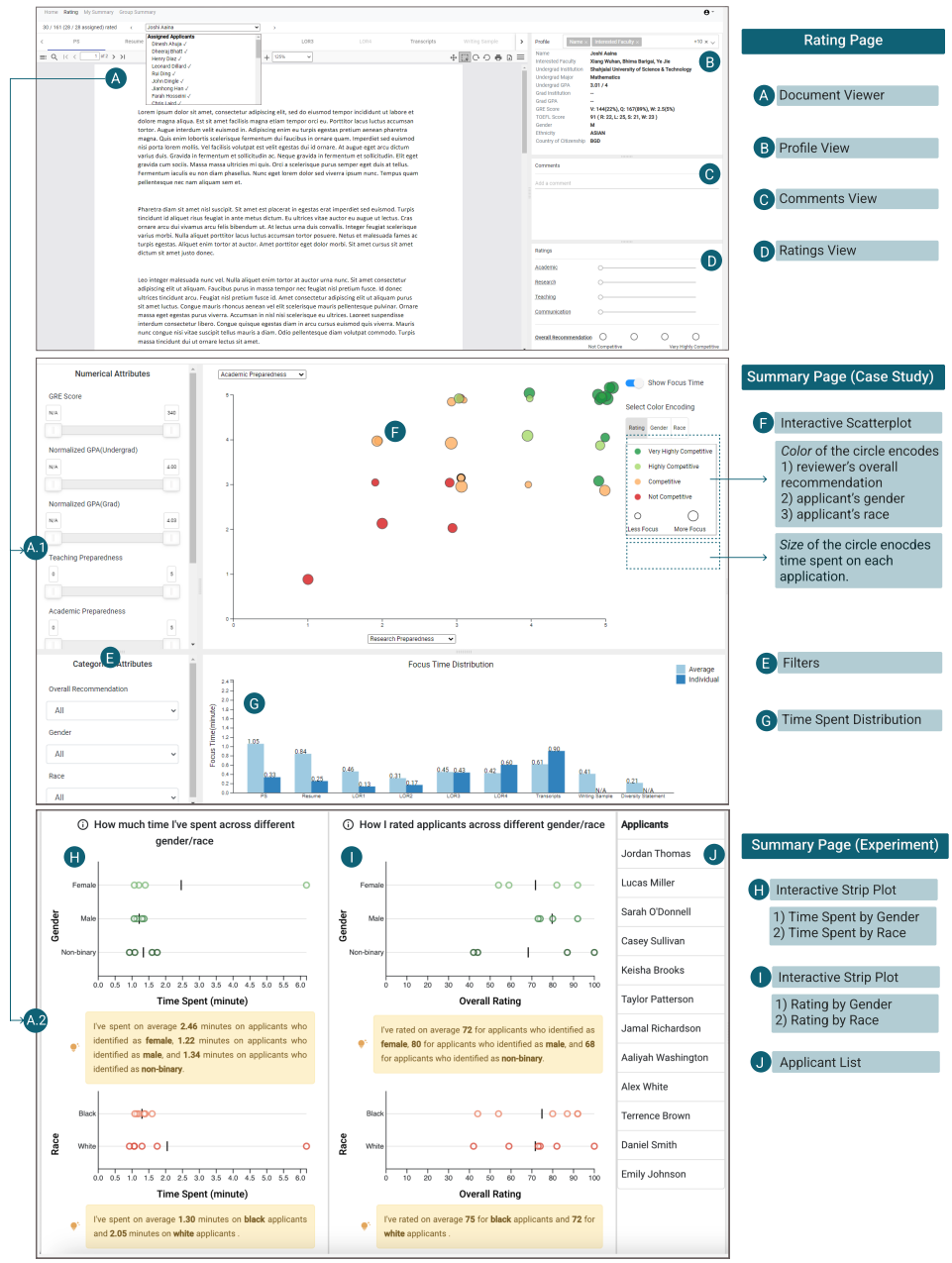

Fig. 1. RUTABAGA supports admissions review with two pages. 1. Rating Page: (A) Documents Viewer shows different application documents, (B) Profile View shows a set of attributes and a drop-down list for selecting/deselecting visible attributes, (C) Comments View and (D) Ratings View. 2. Summary Page (case study, middle) maintains the right-hand-side of the interface for profile, comments, and ratings of individual applicants, and replaces Document Viewer with (E) Filters, (F) Interactive Scatterplot, and (G) Time Spent Distribution Chart. The Summary Page was modified for the controlled study (bottom) which shows strip plots representing users' Time Spent by Gender (H, top) and Race (H, bottom), and Competitiveness Rating by Gender (I, top) and Race (I, bottom) as well as the full applicant list (J). Note that all the application material displayed here is fake data for demonstration purposes.

(described below) with the applicant's data, and *clicking* a point returns to the applicant's portfolio in the Rating Page. Additionally, the scatterplot allows for additional encodings of size (time spent) and color (overall recommendation of an applicant, applicant gender, or applicant race).

The scatterplot is designed based on the need for providing visualizations that allow reviewers to observe patterns and outliers in their time spent and ratings (DG2). While there are many visualization techniques for representing multi-dimension data such as parallel coordinates and scatterplot matrix, we chose a scatterplot because of its effectiveness in identifying patterns/outliers, visual simplicity, and familiarity to general audiences [30] (DG3).

(G) Time Spent Distribution shows a grouped bar chart depicting (i) the reviewer's average time spent on different documents and (ii) the reviewer's distribution of time spent on application components for the hovered applicant in the scatterplot. This view is designed to help users gain insights about the time they spent across different application components (DG2) and identify outliers at the individual applicant level such as little or no review of a certain file for an applicant.

5 CONTROLLED STUDY METHODOLOGY

We conducted a pre-registered[1] crowdsourced study to assess (i) the underlying assumption of RUTABAGA that implicit biases (as measured by implicit association test scores [23]) correlate to observable differences in application review behaviors and decisions and (ii) the effectiveness of the RUTABAGA in enhancing awareness of undesired behaviors. Participants conducted a simulated review of 12 undergraduate university applications using a variation of RUTABAGA described below (Section 5.3).

5.1 Participants

We initially recruited 75 participants through the Prolific crowdsourcing platform based on power analysis (with $\beta = 0.05$, $\alpha = 0.8$) from a pilot study. Seven participants who had invalid gender or race IAT scores (due to rapid responses) were excluded, and we recruited additional participants to maintain our target sample size of 75 (31 identified as female, 40 as male, and 4 as non-binary). Eligibility criteria included 18+ years old, hold a Bachelor's degree or higher, and fluent in English. Twelve participants indicated prior experience in university admissions. Participants spent 77 minutes on average ($min = 25$, $max = 162$) and were compensated $15.

5.2 Dataset

The dataset used in this experiment was comprised of 12 artificially generated university application packets. While we would ideally explore a diverse range of racial and gender identities, we simultaneously were constrained to limit the average study duration to 1 hour to minimize participant fatigue. Therefore, we opted to study three levels for gender (non-binary, female, male), two levels for race (Black, White)[4], and two levels for competitiveness (high, moderate), resulting in 12 unique application portfolios (3x2x2) as summarized in Table 1.Each application packet consisted of a personal statement, resume, transcript, two recommendation letters, and basic form-field information including GPA, SAT score, class rank, gender, and race. We used ChatGPT to generate the application packets using a sequence of prompts shared in Supplemental Materials.

[1] https://aspredicted.org/blind.php?x=N51_KWZ

#	Name	Gender	Race	Competitiveness	SAT Score	GPA	Class Rank
1	Aaliyah Washington	Female	Black	High	1500	3.9	5 / 100
2	Keisha Brooks	Female	Black	Moderate	1290	3.4	90 / 150
3	Emily Johnson	Female	White	High	1550	3.8	14 / 200
4	Sarah O'Donnell	Female	White	Moderate	1280	3.4	163 / 250
5	Jamal Richardson	Male	Black	High	1540	3.9	12 / 200
6	Terrence Brown	Male	Black	Moderate	1260	3.2	68 / 100
7	Lucas Miller	Male	White	High	1490	3.7	14 / 150
8	Daniel Smith	Male	White	Moderate	1300	3.5	180 / 250
9	Taylor Patterson	Non-binary	Black	High	1520	3.8	16 / 200
10	Jordan Thomas	Non-binary	Black	Moderate	1300	3.5	183 / 250
11	Alex White	Non-binary	White	High	1530	4.0	1 / 100
12	Casey Sullivan	Non-binary	White	Moderate	1270	3.3	105 / 150

Table 1. A summary of the 12 artificially generated applications.

5.3 Interface

We created a variation of RUTABAGA for the controlled study which included the Rating page (as described in Section 4.1) and a modified version of the Summary page (as described in Section 4.2). The Summary page (Figure 1, bottom) was updated primarily to adjust features that do not scale down well for the study where only 12 applicants were rated. Namely, we removed the filters (E) and replaced the scatterplot (F) and bar chart (G) with strip plots (H, I) that showed the distribution of users' time spent and overall ratings across different gender/race groups. Each circle on the strip plot represents an applicant and the vertical line indicates the mean value. We added an applicant list (J) to allow easy navigation between applicants with brushing and linking between views. The Ratings panel (D) was also updated to a slider ranging from 1-100 that allowed participants to give an overall competitiveness score for an applicant for a more granular analysis.

5.4 Procedure

This study utilized a within-subjects design where participants first reviewed the applications without the intervention (Control) and were then directed to the Summary Page (Intervention) after rating all the applications. This design allowed us to account for individual differences in review behaviors and decisions.

Participants first completed a step-through tutorial that introduced the system's layout for displaying application portfolios and operations for entering ratings. After completing the walk-through training, participants reviewed 12 applications (order randomized) and rated them on a scale of 1 (not competitive) - 100 (very competitive). After participants rated all the applications, they completed a second step-through tutorial that introduced the features in the modified Summary Page (Figure 1, bottom). Participants were able to interact with the page and revisit profiles to finalize their ratings. The post-study survey included background information such as gender, education level, prior experience in university admissions, and Likert-scale questions about the summary page. Finally, participants completed the Gender-Science and Race implicit association tests (IAT) adapted from Project Implicit [27] (order randomized).

5.5 Hypotheses

We organize our hypotheses into three groups:

H1: Participants' review behavior (time spent) and decisions (competitiveness ratings) will differ based on applicants' race and gender.

- H1.1. Female and non-binary applicants will receive less *review time* than male applicants.
- H1.2. Female and non-binary applicants will be rated as *less competitive* than their equally-qualified male counterparts.
- H1.3. Black applicants will receive less *review time* than White applicants.
- H1.4. Black applicants will be rated as *less competitive* than their equally-qualified White counterparts.

H2: Participants' review behavior (time spent) and decisions (competitiveness ratings) will correlate with their implicit biases.

- H2.1. Participants with higher gender IAT scores will spend *less time* on female and non-binary applicants than male applicants.
- H2.2 Participants with higher gender IAT scores will give *lower ratings* to female and non-binary applicants than male applicants.
- H2.3 Participants with higher race IAT scores will spend *less time* on Black applicants than White applicants.
- H2.4 Participants with higher race IAT scores will give *lower ratings* to Black applicants than White applicants.

H3: Visualizations of the review process will affect reviewers' behavior and decisions.

6 CONTROLLED STUDY RESULTS

In the forthcoming analyses, we normalized participants' time spent and competitiveness ratings to fall between 0 and 1 to control for individual variations. For the analyses of H1 and H2, we used participants' time spent and competitiveness rating data before they interacted with the summary page (Control), while the analysis of H3 considered interactions both before and after the summary page (Intervention).

6.1 H1: Effects of Applicant Gender and Race

We stratified our analysis of time spent and competitiveness ratings by applicant race and gender. Figure 2 summarizes the normalized time spent reviewing applicants (left) and normalized competitiveness ratings (right) for applicants by gender (top) and race (bottom). Each depicts a boxplot representing the 75 participants' time spent or ratings for each applicant that falls into the respective gender or race group. We used a non-parametric test (Mann-Whitney U test) instead of the pre-registered t-test for the analysis since the data do not follow a normal distribution as assessed by the Shapiro-Wilk's test ($p < 0.05$).

We found **partial support for H1.1**. Female applicants ($Mdn = 0.209$) received less review time than male applicants ($Mdn = 0.253$) ($p = 0.004$). On the other hand, non-binary applicants ($Mdn = 0.299$) received more review time than male applicants ($Mdn = 0.253$); however, the difference is not statistically significant ($p = 0.156$). Additionally, we found **no support for H1.2:** female ($Mdn = 0.566$) and non-binary applicants ($Mdn = 0.555$) were rated less competitive than male applicants ($Mdn = 0.633$) ; however, the differences lack statistical significance ($p = 0.656$ and 0.920 respectively). We also found that participants spent more time reviewing Black applicants ($Mdn = 0.261$) than White applicants ($Mdn = 0.245$) and gave them higher average ratings ($Mdn = 0.586$) than White applicants ($Mdn = 0.580$) . These results contradict the directionality of our hypothesis H1.3 and H1.4, but also lack statistical significance ($p = 0.140$ and 0.286 respectively).

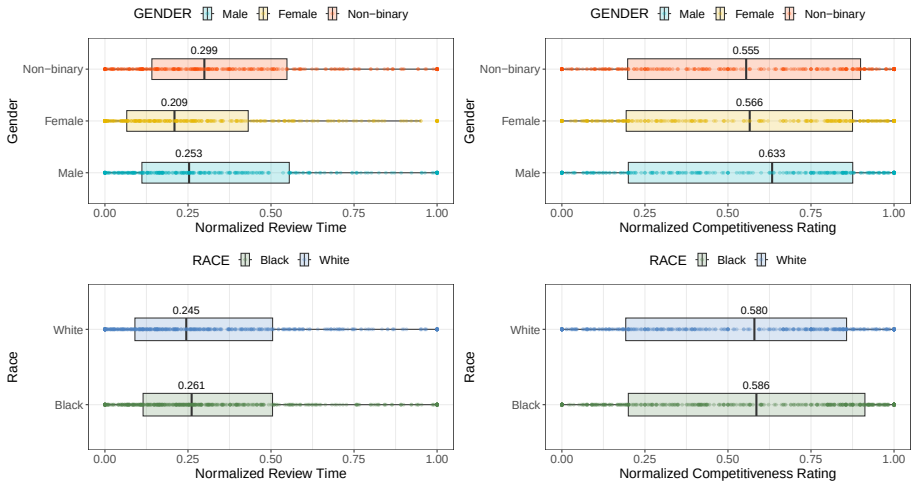

Fig. 2. Boxplots with individual points of normalized review time (left) and normalized competitiveness ratings (right) by applicant gender (top) and race (bottom). The values are normalized per participant, where 0 represents the minimum review time/rating, and 1 represents the maximum review time/rating to facilitate a comparison that accounts for individual differences.

We further validated that findings for H1 were not influenced by order effects. Details of the analysis are included in Supplemental Materials.

6.2 H2: Implicit Biases

To investigate the extent to which implicit biases correspond to deviations in time spent and competitiveness ratings for applicants of different races and genders, we conducted linear regression analyses on the race and gender Implicit Association Test (IAT) scores compared to the disparities in participants' review behaviors and decisions.

The IAT Scores range from -2 to 2, where a higher score indicates a stronger inclination to associate females with liberal arts (Gender IAT) or Black with bad (Race IAT). For each participant, we computed the differences in (i) average time spent on applicants of different gender and race groups and (ii) average competitiveness ratings for applicants of different gender and race groups. We conducted linear regression analyses to predict the gender and race IAT scores using these differences. The core assumptions of the linear regression analysis were checked for each model using residual vs. fitted value plots and normal Q-Q residual plots, and no significant deviations were observed.

We found **no support for H2.1 and H2.1**: a higher gender IAT score (stronger implicit association of females → liberal arts and males → science) did not result in participants spending more time on male applicants than female applicants or giving higher ratings to male applicants compared to female/non-binary applicants ($Coef. < 0$). However, we found **support for H2.3 and H2.4:** participants with higher gender IAT scores spent more time on male applicants than non-binary applicants ($Coef. = 0.044$, $p = 0.019$, corresponding to 0.76 minutes difference between the reviewers with the lowest possible gender IAT score and the highest). Similarly, we observed that participants with higher race IAT scores (stronger implicit association of White → good and Black → bad) indeed spent more time ($Coef. = 0.096$, $p < 0.001$; corresponding to 1.67 minutes difference between the reviewers with lowest possible race IAT score and the highest) and assigned higher ratings ($Coef. = 0.08$, $p < 0.001$; corresponding to 16.92 out of 100 difference in

competitiveness ratings between reviewers with lowest and highest race IAT score) to White applicants than Black applicants.

6.3 H3: Effect of the Intervention

We analyzed participants' interactions on the summary page and compared competitiveness ratings and time spent before and after seeing the summary page. Additionally, we summarize qualitative feedback from participants.

6.3.1 Interaction Analysis. We analyzed logs of click and hover interactions to understand how participants used the interface. *Hover* interactions suggest instances where participants explored their ratings and time spent on applicants. 71 participants (95%) initiated 2192 hover events (1500 from the applicant list and 667 from the strip plots). *Clicking* on points in the strip plots or applicant names indicates revisiting the rating page for the corresponding application. There were 80 revisits from 44 (59%) participants and each of these participants revisited on average 2 applicants (*max* = 7, *min* = 1). Most revisits occurred through the applicant list (53), while about 1/3 of the revisits were initiated from the strip plots (26). Notably, there were more interactions on the strip plots representing Time Spent (456 hovers + 14 clicks) than the strip plots representing Competitiveness Ratings (211 hovers + 12 clicks). This could be attributable to participants' expressed surprise on the Time Spent plots as discussed in the following Section 6.3.3. We found 26 changes to applicant ratings from these revisits (19 upgrades, 7 downgrades) in total for 11 applicants from 9 participants.

We observed certain trends in the types of applicants that participants interacted with: in the Time Spent by Gender and Competitiveness Rating by Gender strip plots, participants interacted with more female applicants compared to non-binary and male applicants; and in the Time Spent by Race and Competitiveness Rating by Race plots, participants interacted with more Black applicants compared with White applicants. The detailed interaction counts are summarized in Supplemental Materials. Furthermore, we found that most of the interactions (10/12 clicks, 144/211 hovers) on the Competitiveness Rating plots corresponded to applicants who were rated lower (below the average). Similarly, most of the click interactions (9/12) on the Time Spent by Gender plot corresponded to applicants who received less review time (less than the average).

6.3.2 Change of Time Spent and Competitiveness Ratings. We compared participants' time spent and competitiveness ratings across different gender/race groups before and after interacting with the summary page and found that the changes are marginal at the aggregate level. However, we found that the summary page had a **substantial impact on some individual participants' review behavior and decisions**. For example, before interacting with the summary page, $P_{CTRL}69$ spent on average more time on White applicants than Black applicants ($\Delta_{\text{mean\_time(W-B)}} = 2.245-1.843 = 0.402$ minutes) and rated White applicants higher than Black applicants ($\Delta_{\text{mean\_rating(W-B)}} = 78.000 - 72.333 = 5.667$). The participant indicated surprise and disappointment about the behavior – *"I was also surprised, and perhaps disappointed with myself, for giving white people higher scores and more time than black people. I was not conscious that I was doing this and it is good to know."* and made adjustments to their evaluation process. As a result, the disparities in time spent and competitiveness ratings among Black and White applicants decreased ($\Delta_{\text{mean\_time(W-B)}} = 2.408 - 2.292 = 0.115$ minutes, $\Delta_{\text{mean\_rating(W-B)}} = 77.667 - 74.167 = 3.50$) after the participant revisited applicants and adjusted competitiveness ratings.

Similarly, participant $P_{CTRL}40$ initially rated male applicants higher than female applicants on average with a substantial disparity ($\Delta_{\text{mean\_rating(M-F)}} = 83.750 - 64.750 = 19$). In response, the participant indicated *"I think I ultimately revised some answers to ensure no disparities apart from objective criteria, in some cases I simply hadn't remembered correctly where I had ranked other students of similar GPA and class rank so I rectified that."*. Consequently, the magnitude

of the disparity decreased significantly, and the direction changed to favor female applicants($\Delta_{\text{mean\_rating(M-F)}}$ = 64.00 − 67.25 = -3.25).

6.3.3 Qualitative Feedback. **Usefulness.** Participants rated the usefulness of the Time Spent (Figure 1 H) and the Competitiveness Rating (Figure 1 I) strip plots on a 5-point scale (1: Not at all useful, 5: Extremely useful). The mean score was 2.9 (SD = 1.1) for the Time Spent plots and 3.2 (SD = 1.2) for the Competitiveness Rating plots.

Many participants (7 for the Time Spent plots, 15 for the Competitiveness Rating plots) commented the plots helped them **assess fairness and potential bias** (e.g., $P_{CTRL}79$ expressed *"It gave me an idea of how much effort I put into each individual and assured me on how fair I treated each individual's information in arriving at my decision"*. Some participants (5) also mentioned that the time spent plots could lead to **adjustment of behavior** (e.g., $P_{CTRL}25$ commented *"I think it could help you reconsider an application if you spent significantly less time on it, and maybe taking another look at an applicant you may have initially dismissed"*). Some participants (3) commented that the Competitiveness Rating plots would be *"more useful if there was a larger data pool"* ($P_{CTRL}68$) suggesting that the visualizations could have more impact in real-world admissions scenarios where the application pool is much larger.

Participants who found the visualizations not useful emphasized that they were fair – their evaluations were based on applicants' merit without considering gender or race. Some participants thought the time spent plots were not useful because the time spent on applications was influenced by different factors (*"I didn't find it particularly useful just because there were other factors at play."* - $P_{CTRL}8$) including the order in which they were reviewed and the content of the application package.

Surprise. Participants rated whether they were surprised by their data shown on the Time Spent plots (Figure 1 H) and the Competitiveness Rating plots (Figure 1 I) on a 5-point scale (1: Not at all surprised, 5: Extremely surprised). In general, participants expressed no/low surprise in their data with a mean score of 1.9 (SD = 1.0) for Time Spent plots and 1.7 (SD = 1.0) for Competitiveness Rating plots. Many participants who were not surprised mentioned that gender and race were not factors that impacted their reviews. In some cases, participants expressed surprise that they rated applicants in a certain group higher (e.g., *"I was surprised that I tended to rate male applicants higher than female applicants."* - $P_{CTRL}44$) or spent more time on a certain group of applicants unconsciously (e.g., *"I wasn't aware that I spent more time reviewing non-binary applicants."* -$P_{CTRL}12$).

7 CASE STUDY METHODOLOGY

We conducted a case study in the Computer Science department at a private university where the system was used for the department's Ph.D. admissions reviews over a time period of roughly two months. This case study allowed us to assess real-world efficacy of RUTABAGA for bias awareness.

7.1 Participants

We recruited participants among the faculty in the Computer Science Department at a large private university. The admissions committee chair sent out an email to all admissions committee members with an introduction and training materials (documentation and video tutorial) of RUTABAGA. The committee members were recommended (but not required) to use RUTABAGA; they could still elect to use prior application review methods (Section 3.2). The two committee chairs and 11/12 committee members *used the system* in some capacity during the admissions process. We were able to subsequently *interview* the two committee chairs and nine committee members after the admissions process

concluded. They had 0 to 14 (μ = 4) years of prior involvement in admissions, offering a diverse range of perspectives. We refer to participants in the "Results" section (Section 8) as P1-13. Among these, P1 - P11 used the system and were interviewed, P12 - P13 were not interviewed, and P9 - P11 only used the system partially. We note that one of the authors was a member of the admissions committee (P3). We include their data in our analysis in order to present a comprehensive and transparent view of the admissions review process. Except for P3 and the two committee chairs, none of the participants had prior experience using the system before the case study.

Given the complexity of deploying new tools in real-world scenarios, our study focused on a single department to ensure a thorough evaluation. Although the sample size is limited, it includes experienced decision-makers directly involved in the admissions process, capturing key decision-making dynamics – with committee chairs overseeing the process and faculty members evaluating applications. The study allowed us to gather detailed feedback on the system's usability, effectiveness, and potential to promote bias awareness. The insights gathered, while specific to a single department, are valuable for understanding the system's potential in similar contexts. We discuss the limitations of relying on a convenience sample and its potential impact on the generalizability of the system design in Section 9. Future work will aim to address these limitations by expanding the participant pool to include a wider range of programs and institutions, ensuring that the system is robust and adaptable to various admissions contexts.

7.2 Dataset

There were 161 Ph.D. applications in total, and each application was reviewed by at least two committee members. Each application consists of general form field information including research interests, education background, and optional test scores; required document uploads, including personal statement, resume, up to four recommendation letters, and transcripts; and optional files including writing samples and diversity statement. The applications were downloaded from the application portal and loaded to the database for RUTABAGA by the first author before the system was available for reviewers.

7.3 Procedure

During the first two weeks of the admissions review phase, participants independently reviewed and rated assigned applications. Next, the first author loaded the scores and notes provided in a spreadsheet from faculty members who did not use the system, then enabled the Group Summary Page before the first committee meeting. During the meeting, the Group Summary Page (described in Supplemental Materials) was used to facilitate the discussion via screensharing on Zoom. The committee decided which candidates were not a good fit for the program (assigned to the Reject list) and the rest of the candidates advanced to the Interview round which occurred over the subsequent two-week period. A second committee meeting took place after the interviews where the final admissions decisions were made.

After the admissions cycle concluded, emails were sent to all the committee members to invite them for an interview in order to gather user feedback. All the interviews were conducted via zoom and lasted 30 to 60 minutes. Upon providing informed consent, the interviews were screen-recorded. After gathering background information, the interviewer asked the participant to login to the system and share their screen to facilitate a walk-through of the system, discuss their experience, and provide suggestions for improvements. Following the interview, the participants were asked to complete a post-study questionnaire, consisting of questions about the usability of the system (Section 8.4).

7.4 Analysis and Coding

The first author first used an audio-to-text tool to transcribe the interview recordings, and then manually refined the transcripts. The research team used qualitative data analysis methods[33] to analyze the interview transcripts. Specifically, thematic analysis was conducted on the interviews through inductive coding. Two authors independently coded two interview transcripts and discussed to develop a codebook. After refining coding definitions together, the first author coded the remaining transcripts. The final codebook contains 36 codes in nine categories including System Usage, Review Strategy, Awareness, and others, the details of which are included in Supplementary Materials.

8 CASE STUDY RESULTS

We organize results according to high-level themes of the case study, including increasing participants' **awareness** of their review process (Section 8.1), associated **behavioral** changes (Section 8.2) and changes in **decisions** (Section 8.3), consistent with [53]. System **usability** scores and qualitative feedback are presented in Section 8.4.

While 11 reviewers used the Rating Page, only five of them used the Summary Page, although all provided feedback after interacting with the view during the interview. All of the reviewers interacted with the Group Summary Page directly or indirectly (through screen-share from the committee chair) during the committee meetings.

8.1 Awareness

We use the term **awareness** to refer to insights gained from reflecting on one's review process via interaction history on the applications. The findings in this section were derived from the interviews we conducted with the participants.

Process Awareness. Three participants found that features in the Summary Page helped them to systematically reflect on fairness in their time spent across applications and application components (*"I was able to look at the amount of time that I spent over their different documents just to make sure I didn't miss anything."* - P8) and adjust their review behaviors when they identified undesired behaviors. For instance, the time spent distribution chart helped P8 be **aware** that (*"I hadn't really spent time on their writing sample"*), led them to different **behaviors** (*"So I went back and I had the chance to read over it"*), and make changes to some **decisions** (*"it was a good chance to revise what I had scored for their research preparedness"*). The scatterplot helped P3 identify outliers in the time they spent on applications (*"I just didn't spend that much time on someone"*) and led to different **behaviors** (*"I would try to go back and (...) just spend a little bit more time looking at them and see if I missed something"*). P5 commented, *"the folks that I may have not spent as much time on, I'd want to know why."*

Outcome Awareness. In addition to assessing internal consistency of time spent, three participants also used the Summary Page to check internal consistency in their ratings (*"I check the different distributions just to make sure that I was sort of consistent in giving my overall ranking."* -P8. P8 tried to self-calibrate on the ratings (*"we had the chance to go back to our reviews, compare the ones that we had scored previously, and change our rating"*). The scatterplot helped P3 identify outliers in the ratings (*"I identified outliers like someone that I rated as having high research preparedness but I did not rate them overall very highly."*) and revisited the applications (change in **behaviors**) – *"I would go back and look at their applications again"*. The committee chairs planned to *"take a look at the scatterplot to evaluate our process overall"* in the Group Summary Page.

Fairness/Bias. Four participants found the system useful in terms of increasing awareness about procedural fairness. P5 liked that the profile view allows hiding attributes (*"I really like that... I basically turned off anything that I felt might*

bias my decision"). P8 was interested in seeing *"was there really some sort of unintentional way of aspects that influence my decision."* By looking at the scatterplot in the Summary Page with different combinations of X- and Y-axis attributes, the participant found that *"there's no bias towards any gender or race in my decision. That was good for me to know."*

During the interview, participants who did not actively use the Summary Page during the admission process tried to interact with the interface and found the scatterplot *"is showing how I have reviewed people (...) if I have some gender bias or race bias."* (P4), and the time spent chart could answer the question *"are you spending the right amount of time or at least enough time on all the different applicants?"* (P2), indicating the system has the potential to increase reviewers' awareness of bias during the review process, even if they did not initially use the system as such.

8.2 Interaction Analysis

Interactions with Individual Summary. The system logged users' interactions with the Summary Page, including hovering on the points on the scatterplot (Figure 1F) (to see the applicant's information), clicking on the points (to revisit the Rating Page), and clicking on the Overall Recommendation radio buttons (to modify overall recommendation). We analyzed this interaction data to understand if and how participants used this part of the interface. We found that reviewers visited the Summary Page at different phases of the admission process, i.e., during the individual review phase and during the group meeting. Of the 11 reviewers who used the Rating page, five visited the Summary Page. Among these, three actively interacted with the page during both phases and the other two reviewers interacted with the page only during the group meeting. Table 2 shows the number of distinct applicants these five reviewers interacted with in the two phases. Although the numbers are too small to draw generalizable conclusions, we further looked into what type of applicants reviewers tended to hover and click on. During the independent review phase, P8 hovered on almost all the applicants they reviewed, revisited applicants with low overall recommendation, and rated the applicants higher after revisits. P1 mostly hovered on applicants who they rated as Not Competitive and rated one of them higher afterward. P3 hovered on and revisited more competitive applicants and both downgraded and upgraded some applicants.

Table 2. Participants' interactions with the Summary Page. Numbers in parentheses represent interactions during the independent review phase.

	# Hover	# Revisit	# Changes in Recommendation
P1	22 (5)	4 (1)	0 (1)
P3	4 (13)	1 (3)	0 (2)
P5	5	0	0
P8	10 (27)	8 (3)	0 (3)
P12	11	1	0

Time Spent. One-way ANOVA shows that at the aggregate level, there were no significant differences in the time spent across gender and race. To account for differences in application characteristics, we normalized the time spent by incorporating a complexity score for each application. This complexity score was calculated as the sum of the normalized portfolio length and normalized word count. By normalizing time spent, we ensure that comparisons reflect differences in reviewer behavior rather than variability in the complexity of the applications themselves. However, similar to our findings from the controlled study, we observed individual differences on how reviewers spent time across gender. As shown in Figure 3, some reviewers (e.g., P3 and P4) spent more time on Female applicants on average, while others (e.g., P2 and P12) spent more time on Male applicants. Among the 10 participants in Figure 3, two identified as female (P3 and P5), while eight identified as male. Of the female participants, one spent more time reviewing female applicants,

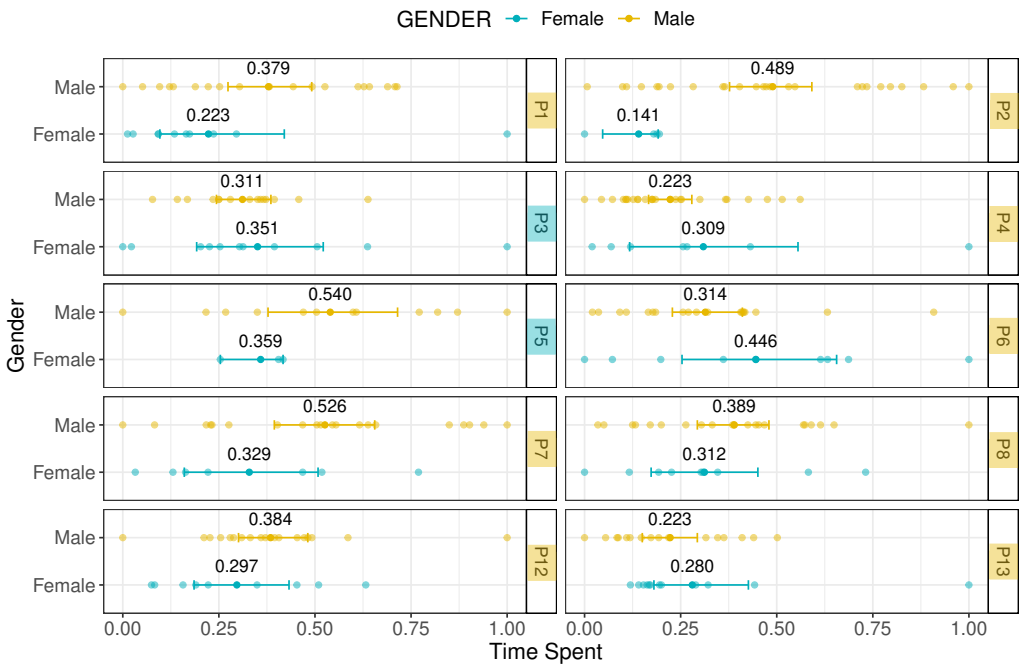

Fig. 3. Strip plots comparing time spent (normalized) on applicants by each reviewer grouped by applicant gender with bootstrapped 95% CI. The mean value is annotated for each gender and each participant. The background color of each participant label represents the reviewer's gender: blue for female participants and yellow for male participants. Three participants' data were excluded due to limited usage of the system for application review.

while the other spent more time reviewing male applicants. The review behavior among the male participants was also variable. Some (e.g., P4 and P6) spent more time reviewing female applicants, some (e.g., P2 and P7) spent more time reviewing male applicants. Given the small sample size, it is not feasible to identify a consistent trend in review behavior linked to reviewer gender.

Those who interacted with the awareness features (P1, P3, P8) tended to have relatively small disparities in time spent between male and female applicants (P1, P8) or, in some cases, reversed the trend to spend more time on female applicants (P3). We note that although many comparisons did not result in statistically significant differences at the group level, the analyses of individual reviewer behavior can be cause for further scrutiny. We discuss this, along with potential explanations of these trends in Section 9.

8.3 Decisions

The distribution of admitted, waitlisted and rejected applicants by gender and race aligns to the underlying distribution of the candidate pool, i.e., there was no clear favor of a certain group when making decisions (Figure 4), confirmed by a Chi-square test ($p > 0.1$). As described in Section 8.2, reviewers' individual recommendations for some applicants were altered after interacting with the Summary page. There were four upgrades and two downgrades from 3 reviewers on 6 applicants. Two of the applicants whose ratings were upgraded were admitted while the two downgraded applicants

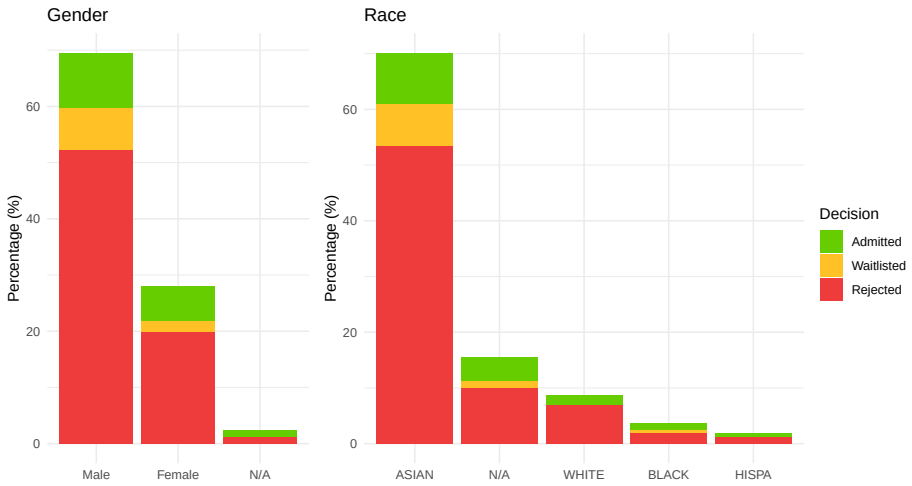

Fig. 4. The distribution of admissions decisions across applicant gender and race.

were rejected, suggesting that the intervention could have had a critical impact on the outcomes for these applicants, although we cannot say with certainty what the outcome would have been for these applicants otherwise.

8.4 System Usability

In the post-study questionnaire, participants rated their overall experience with the system as well as impressions on different features (details are attached in Supplemental Materials). Results of the questionnaire and qualitative feedback from the interviews indicate that participants' overall experience with RUTABAGA was positive.

Participants' average system usability scale (SUS) score is "Good" (score above 68) at 78.3 [7]. Visualization components of the system (the Scatterplot and the Time Spent Distribution) each received divergent usefulness scores ranging from 1 (low) to 5 (high). The mean score was 3.4 ($SD = 1.1$) for the Scatterplot and 3.5 ($SD = 1.1$) for the Time Spent Distribution. P11 commented that the scatterplot can be useful for identifying patterns and suggested providing guidance on how to perform analyses with it. Participants who rated the Scatterplot relatively low commented that *"faculty are (...) very busy... So we don't necessarily spend maybe enough time being self critical"* and suggested that the analysis can be done by administrators (P2). We discuss future work on incorporating these suggestions in Section 9.

9 DISCUSSION

Implications of Affirmative Action. Affirmative action was a policy that arose in the United States in the 1960s to reinforce the Civil Rights Act of 1964 aimed at eliminating discrimination [3]. It intended to create opportunities for traditionally disadvantaged and underrepresented groups in workplace and educational settings and thus, alongside standard evaluation criteria, involved explicit consideration of additional factors such as race and gender in application review processes [18].

The 2023 Supreme Court decisions (SFFA v. Harvard and SFFA v. UNC) ruled against considering race as an explicit factor of admissions decisions [38]. Many admissions committees are operationalizing the ruling by removing direct access to racial identity data *during* the admissions process. Thus programs may choose to restrict real-time access to visualizations shown in the Summary page (1, middle) to instead only allow for retrospective analysis to ensure

compliance. Nevertheless, the Supreme Court maintains that race may be considered insofar as it affects an individual's experiences. To maintain diverse student bodies within these legal constraints, universities have adopted alternative recruiting strategies. Many have expanded outreach efforts to first-generation students, low-income students, and students from rural settings, enhanced financial aid programs, and introduced new essay prompts to allow applicants to discuss their life experiences and challenges [8, 25]. Despite these efforts, the ruling has led to notable changes in the demographic composition of new classes at many top institutions, with black student enrollment declining [8, 25]. Navigating the legality of ensuring diversity, equity, and inclusion in university admissions remains an ongoing challenge in the wake of SFFA v. Harvard and SFFA v. UNC rulings. As a result, we maintain that the importance of continued research on ensuring fair review processes has never been more critical.

In the future, we can explore methods to assist admissions committees in meaningfully accounting for the differences in opportunities and privileges among diverse applicants. For instance, new system features could be developed to help reviewers identify various forms of adversity faced by applicants and design targeted interventions to ensure that these students are not overlooked in the admissions process.

Is Time Spent a Good Proxy for Bias? More time spent on an applicant does not necessarily reflect a negative bias. If a reviewer spends less time on a specific applicant, it could be due to unconscious bias, but it could also be due to other benign reasons. For instance, P11 noted *"I just know this candidate and I wrote her letter, and so I'm looking very little at her."* Other reviewers observed that *"just like in a conference reviewing setting, there are some manuscripts that are clear accepts and there's some manuscripts that are clear rejects."* -P2. Other factors also influenced time spent such as a reviewer's familiarity with transcripts from foreign institutions (P6), general readability of other application components (P6), or varying lengths of documents like recommendation letters (P5). Reviewers tended to agree that *"most of the time is being spent on the murky middle"* -P11. Participants from the controlled study also indicated that they tended to naturally spend more time on the first few applicants when getting familiar with the task.

Thus time spent on applications is an inherently noisy metric sensitive to several factors, rather than a definitive measure of bias. However, results of our controlled study in Section 6.2 indicate that **in spite of these sources of noise, time spent can nonetheless provide a meaningful signal** that illustrates an observable effect of implicit bias on review behaviors. In addition, similar to the stated goals for Wall et al's bias metrics [51] and consistent with the goals of reflective design [44], our aim in RUTABAGA is to promote *reflection* on potential biases. Thus while time spent is an imperfect proxy for bias, its representation in RUTABAGA can cause reviewers to more carefully reflect on their review process nonetheless. Future work can explore variability in application characteristics, such as readability, to better understand the factors influencing time spent on applications. Incorporating these measures could enhance the interpretation of time spent as a signal for bias and provide deeper insights into review behaviors.

Design Implications. While most of the participants from the controlled study interacted with the features intended for reflection to some extent (Section 6.3), only part of the committee used the features in the case study (Section 8). Despite close collaboration with committee chairs during the system's formative design (Section 3), many committee members nonetheless found the Summary and Group Summary pages to be visually overwhelming and chose not to engage (*"It looks scary to me ... I felt a little bit overwhelmed by what was going on"* -P11). Given our observations from the post-study interviews in the case study, we iterated on the system design to enhance user engagement by making the system more intuitive and accessible. The changes include: 1) adding an embedded Help Page in the system for a more convenient on-boarding, 2) providing an explanation of the Summary page when the user first interacts with the page, 3) adding a text summary for each of the visualizations to reduce cognitive load, and 4) adding pre-defined

box-plots to show the user's time spent across different gender/race groups. We observed increased usage of the system in the subsequent admission cycles (27 admissions committee members engaged with the Summary page).

Based on reflection of lessons learned from our studies, we summarize some design considerations for future work creating visual tools for admissions review.

- Augmenting visualizations with text could ease cognitive load. As noted by participants from the controlled study, text summaries of the visualizations (which were absent in the case study) *"helped to ease the processing of the data"* - $P_{CTRL}43$.
- Providing pre-defined analyses could complement self-exploration. In the case study, participants suggested pre-defined analyses, e.g., committee chairs configuring views of essential bias analyses such as gender distribution of admitted applicants or time spent by applicant gender individual committee members can start with.
- Providing sufficient onboarding is critical for system adoption. We provided a tutorial video demonstrating the features of the system in the case study. However, some participants didn't know how to use it and *"needed a better training"* (P7). Integrated explanatory features such as clickthrough tutorials, embedded videos, and help pages could increase learnability.
- Exploring the balance of mixed-initiative user interfaces [26, 54] could be a promising direction. The system could gently nudge [48] users to interact with reflective views or, more aggressively in cases of strict procedural goals, require it, e.g., using pop-up notifications that cannot be dismissed prior to engagement with the analysis.

Impact of the Intervention on Individuals. We observed from the studies that (i) individual behaviors can be highly variable, and (ii) the impact of the awareness features of RUTABAGA is likewise variable. The efficacy of techniques used in RUTABAGA is likely sensitive to reviewers' conscientiousness, time constraints, and other individual factors; some individuals may be resistant to any form of intervention in this setting. However, we are optimistic that our analysis of individual reviewers revealed that RUTABAGA's techniques have the potential to be transformational for some reviewers' processes. The system has demonstrated the potential to cause reflection and subsequent changes in behaviors and decisions for some individuals who are motivated to assess and rectify behavioral disparities.

Generalizability and Scalability. The current system is designed based on one program's Ph.D. admissions process. However, the admissions process can vary widely across different institutions and programs at the graduate and undergraduate levels. For example, Master's admissions processes often prioritize academic performance, such as GPA and standardized test scores, while non-academic factors such as personal statements and letters of recommendation play a less significant role in the review [56]. Undergraduate admissions, on the other hand, typically involve larger applicant pools and often rely on a broader range of holistic evaluation criteria that include factors such as extracurricular activities, socioeconomic status, and academic records [12]. Nevertheless, our design could be adapted to support holistic review processes similar to ours. The specific features like the rating criteria and scales, plot configurations, etc. in the system can be customized for different program needs. However, further work is needed to assess its applicability to programs deviating substantially from ours, e.g., in size of application pool, distribution of reviewer responsibilities, etc. We could envision additional features that may be helpful in the case of large applicant pools. For instance, boxplots or density plots can be used instead of strip plots to present high-level summaries with outliers displayed. Users can observe aggregated trends first and dive into individual-level data as desired. For institutions already employing customized tools (e.g., Slate [1]) for admissions review, capturing user interactions and providing visualizations could

be integrated into existing tools by closely working with the tool designers. Seamless integration, however, can be challenging and needs to be explored in future work.

Ethical Considerations in Real-World Deployment. Deploying the system in a real-world admissions context introduces ethical considerations, particularly given its potential to influence applicant outcomes. The deployment of RUTABAGA was guided by the department's goal of improving fairness in the admission process. The system was designed to address the inherent limitations in human decision-making by identifying potential biases and inconsistencies. Its role was not to replace human reviewers, but rather to act as an assistive tool that could identify issues that might otherwise have gone unnoticed. To mitigate potential concerns, admissions committee members were fully informed about the purpose and capabilities of the system, and the use of the system was entirely optional for reviewers.

Limitations. One limitation of our work is that participants in the controlled study were not screened for admissions training. As we discussed in Section 5.1, it is challenging to recruit a substantial sample size of admissions officers, and instead we recruited participants based on a minimum education level of a bachelor's degree as inclusion criteria. Admissions reviewers typically receive diversity, equity, and inclusion (DEI) training, which may influence their decision-making differently from untrained participants. Nevertheless, our findings may be comparable to what could be expected in general workplace application reviews, where the background and training of reviewers tend to be less standardized and more diverse than in Ph.D. admissions [55]. Nevertheless, even in our real-world deployment involving trained Ph.D. admissions reviewers, we observed high variability in review behaviors, underscoring the complexity of decision-making in these contexts. This variability suggests that individual differences play a significant role in reviewer behaviors, regardless of training or experience. However, we cannot speculate on how the findings of this controlled study would generalize to larger sample sizes or other domains without further empirical investigation.

Another limitation of the control study is the limited set of gender and racial categories used. Although our goal was to obtain comparable results with previous work [4] and to create as diverse a range of applicant profiles as possible, we were constrained by the need to limit the study duration to prevent participant fatigue. The exclusion of other racial groups, such as Hispanic or Asian applicants, constrains our ability to explore the broader dynamics of bias and fairness. Different racial categories may reveal unique patterns of reviewer behavior or bias that were not captured here. Future research should incorporate a more comprehensive range of racial categories to better understand how biases operate across diverse populations.

Furthermore, the interviews conducted in the case study may be subject to response bias [19], where participants may provide feedback they perceive as desirable and favorable by the researchers.

10 CONCLUSION

In this paper, we presented RUTABAGA, a system designed to facilitate bias-aware admissions decision making. Designed alongside two graduate program admissions committee chairs, the system allows the admissions committee members to independently review applications, reflect on their review process, and collaboratively discuss, calibrate, and make admissions decisions. Reviewers' interactions with applications are recorded to capture time spent across applicants and application components and visualized to promote self-reflection of the review processes and increase awareness of potentially biased processes. We examined and confirmed the underlying assumption of the RUTABAGA system in a controlled study that demonstrates implicit racial bias correlates to observable differences in review behaviors and decisions. We also evaluated RUTABAGA via a case study in the Computer Science department at a private university

where the system was used for the department's 2022 graduate admissions cycle. We conclude that RUTABAGA is a promising approach to increase awareness and affect changes in behaviors and decisions for individual admissions reviewers.

REFERENCES

[1] [n. d.]. Slate. Retrieved November 30, 2023 from https://slate.org/

[2] Chadia Abras, Diane Maloney-Krichmar, and Jenny Preece. 2004. User-Centered Design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.

[3] Civil Rights Act. 1964. Civil Rights Act of 1964. *Title VII, Equal Employment Opportunities* (1964).

[4] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American economic review* 94, 4 (2004), 991–1013. https://doi.org/10.1257/0002828042002561

[5] Katerina Bezrukova, Karen A. Jehn, and Chester S. Spell. 2012. Reviewing Diversity Training: Where We Have Been and Where We Should Go. *Academy of Management Learning amp; Education* 11, 2 (June 2012), 207–227. https://doi.org/10.5465/amle.2008.0090

[6] David Borland, Jonathan Zhang, Smiti Kaul, and David Gotz. 2020. Selection-Bias-Corrected Visualization via Dynamic Reweighting. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1481–1491. https://doi.org/10.1109/tvcg.2020.3030455

[7] John Brooke. 2013. SUS: A Retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[8] Hilary Burns and Neena Hagen. 2024. 'It is heartbreaking': Black first-year enrollment plunges at top colleges, post-affirmative action. *The Boston Globe* (25 September 2024). https://www.bostonglobe.com/2024/09/25/metro/affirmative-action-colleges-black-enrollment/

[9] Quinn Capers IV, Daniel Clinchot, Leon McDougle, and Anthony G Greenwald. 2017. Implicit Racial Bias in Medical School Admissions. *Academic Medicine* 92, 3 (2017), 365–369. https://doi.org/10.1097/acm.0000000000001388

[10] Mengyu Chen and Emily Wall. 2022. Perception of Skill in Visual Problem Solving: An Analysis of Interactive Behaviors, Personality Traits, and the Dunning-Kruger Effect. (2022).

[11] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The Anchoring Effect in Decision-Making with Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 116–126. https://doi.org/10.1109/vast.2017.8585665

[12] Arthur L Coleman and Jamie Lewis Keith. 2018. Understanding Holistic Review in Higher Education Admissions. *New York: College Board* (2018).

[13] Evanthia Dimara, Gilles Bailly, Anastasia Bezerianos, and Steven Franconeri. 2018. Mitigating the Attraction Effect with Visualizations. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 850–860. https://doi.org/10.1109/tvcg.2018.2865233

[14] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2016. The Attraction Effect in Information Visualization. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 471–480. https://doi.org/10.1109/tvcg.2016.2598594

[15] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2018. A Task-Based Taxonomy of Cognitive Biases for Information Visualization. *IEEE transactions on visualization and computer graphics* 26, 2 (2018), 1413–1432. https://doi.org/10.1109/tvcg.2018.2872577

[16] Frank Dobbin and Alexandra Kalev. 2018. Why Doesn't Diversity Training Work? The Challenge for Industry and Academia. *Anthropology Now* 10, 2 (May 2018), 48–55. https://doi.org/10.1080/19428200.2018.1493182

[17] Mi Feng, Evan Peck, and Lane Harrison. 2018. Patterns and Pace: Quantifying Diverse Exploration Behavior with Visualizations on the Web. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 501–511. https://doi.org/10.1109/tvcg.2018.2865117

[18] Robert Fullinwider. 2018. Affirmative Action. In *The Stanford Encyclopedia of Philosophy* (Summer 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[19] Adrian Furnham. 1986. Response Bias, Social Desirability and Dissimulation. *Personality and individual differences* 7, 3 (1986), 385–400. https://doi.org/10.1016/0191-8869(86)90014-0

[20] David Gotz, Shun Sun, and Nan Cao. 2016. Adaptive Contextualization: Combating Bias During High-Dimensional Visualization and Data Selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 85–95. https://doi.org/10.1145/2856767.2856779

[21] Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological review* 102, 1 (1995), 4. https://doi.org/10.1037//0033-295x.102.1.4

[22] Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit Bias: Scientific Foundations. *California law review* 94, 4 (2006), 945–967. https://doi.org/10.2307/20439056

[23] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of personality and social psychology* 74, 6 (1998), 1464. https://doi.org/10.1037/0022-3514.74.6.1464

[24] Sunwoo Ha, Shayan Monadjemi, Roman Garnett, and Alvitta Ottley. 2022. A Unified Comparison of User Modeling Techniques for Predicting Data Interaction and Detecting Exploration Bias. *IEEE Transactions on Visualization and Computer Graphics* (2022). https://doi.org/10.1109/tvcg.2022.3209476

[25] Anemona Hartocollis. 2024. Harvard's Black Student Enrollment Dips After Affirmative Action Ends. *The New York Times* (11 September 2024). https://www.nytimes.com/2024/09/11/us/harvard-affirmative-action-diversity-admissions.html

[26] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166. https://doi.org/10.1145/302979.303030

[27] Project Implicit. 2023. Implicit Association Test. Retrieved September 13, 2023 from https://implicit.harvard.edu/implicit/selectatest.html

[28] Sarah M. Jackson, Amy L. Hillard, and Tamera R. Schneider. 2014. Using Implicit Bias Training to Improve Attitudes toward Women in STEM. *Social Psychology of Education* 17, 3 (May 2014), 419–438. https://doi.org/10.1007/s11218-014-9259-5

[29] Po-Ming Law and Rahul C Basole. 2018. Designing Breadth-Oriented Data Exploration for Mitigating Cognitive Biases. In *Cognitive Biases in Visualizations*. Springer, 149–159. https://doi.org/10.1007/978-3-319-95831-6_11

[30] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2016. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 551–560. https://doi.org/10.1109/tvcg.2016.2598920

[31] Qianyu Liu, Haoran Jiang, Zihao Pan, Qiushi Han, Zhenhui Peng, and Quan Li. 2024. BiasEye: A Bias-Aware Real-time Interactive Material Screening System for Impartial Candidate Assessment. *arXiv preprint arXiv:2402.09148* (2024). https://doi.org/10.48550/arxiv.2402.09148

[32] Ronald A Metoyer, Tee Chuanromanee, Gina M Girgis, Qiyu Zhi, and Eleanor C Kinyon. 2020. Supporting Storytelling With Evidence in Holistic Review Processes: A Participatory Design Approach. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–24. https://doi.org/10.1145/3392870

[33] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2018. *Qualitative Data Analysis: A Methods Sourcebook*. Sage publications.

[34] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science Faculty's Subtle Gender Biases Favor Male Students. *Proceedings of the national academy of sciences* 109, 41 (2012), 16474–16479. https://doi.org/10.1073/pnas.1211286109

[35] Atilla Alpay Nalcaci, Dilara Girgin, Semih Balki, Fatih Talay, Hasan Alp Boz, and Selim Balcisoy. 2019. Detection of Confirmation and Distinction Biases in Visual Analytics Systems. In *TrustVis@ EuroVis*. 13–17. https://doi.org/10.2312/trvis.20191185

[36] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2021. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1009–1018. https://doi.org/10.1109/tvcg.2021.3114827

[37] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. 2011. Analytic Provenance: Process+interaction+insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 33–36. https://doi.org/10.1145/1979742.1979570

[38] Supreme Court of the United States. 2023. Students for Fair Admissions v. Harvard. Retrieved September 13, 2023 from https://www.supremecourt.gov/opinions/22pdf/20-1199_hgdj.pdf

[39] Brandy Pieper and Masha Krsmanovic. 2022. Faculty perceptions on (implicit) bias during the graduate admission review process. *Studies in Graduate and Postdoctoral Education* 14, 2 (Nov. 2022), 117–133. https://doi.org/10.1108/sgpe-05-2022-0040

[40] Julie Posselt, Theresa E. Hernandez, Cynthia D. Villarreal, Aireale J. Rodgers, and Lauren N. Irwin. 2020. *Evaluation and Decision Making in Higher Education: Toward Equitable Repertoires of Faculty Practice*. Springer International Publishing, 453–515. https://doi.org/10.1007/978-3-030-31365-4_8

[41] Daniel L Reinholz, Samantha Ridgway, Poorna Talkad Sukumar, and Niral Shah. 2023. Visualizing Inequity: How STEM Educators Interpret Data Visualizations to Make Judgments about Racial Inequity. *SN Social Sciences* 3, 5 (2023), 76. https://doi.org/10.1007/s43545-023-00664-0

[42] Daniel L Reinholz and Niral Shah. 2018. Equity Analytics: A Methodological Approach for Quantifying Participation Patterns in Mathematics Classroom Discourse. *Journal for Research in Mathematics Education* 49, 2 (2018), 140–177. https://doi.org/10.5951/jresematheduc.49.2.0140

[43] Daniel L Reinholz, Amelia Stone-Johnstone, and Niral Shah. 2020. Walking the Walk: Using Classroom Analytics to Support Instructors to Address Implicit Bias in Teaching. *International Journal for Academic Development* 25, 3 (2020), 259–272. https://doi.org/10.1080/1360144x.2019.1692211

[44] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective Design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58. https://doi.org/10.1145/1094562.1094569

[45] Cheryl Staats, Kelly Capatosto, Lena Tenney, and Sarah Mamo. 2017. State of the Science: Implicit Bias Review 2017. *Kirwan Institute for the Study of Race and Ethnicity* (2017).

[46] Poorna Talkad Sukumar and Ronald Metoyer. 2018. A Visualization Approach to Addressing Reviewer Bias in Holistic College Admissions. In *Cognitive Biases in Visualizations*. Springer, 161–175. https://doi.org/10.1007/978-3-319-95831-6_12

[47] Poorna Talkad Sukumar, Ronald Metoyer, and Shuai He. 2018. Making a Pecan Pie: Understanding and Supporting The Holistic Review Process in Admissions. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22. https://doi.org/10.1145/3274438

[48] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.

[49] Andre Calero Valdez, Martina Ziefle, and Michael Sedlmair. 2017. Priming and Anchoring Effects in Visualization. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 584–594. https://doi.org/10.1109/tvcg.2017.2744138

[50] Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert. 2019. A Formative Study of Interactive Bias Metrics in Visual Analytics Using Anchoring Bias. In *IFIP Conference on Human-Computer Interaction*. Springer, 555–575. https://doi.org/10.1007/978-3-030-29384-0_34

[51] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115. https://doi.org/10.1109/vast.2017.8585669

[52] Emily Wall, Leslie M Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert. 2018. Four Perspectives on Human Bias in Visual Analytics. In *Cognitive biases in visualizations*. Springer, 29–42. https://doi.org/10.1007/978-3-319-95831-6_3

[53] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2021. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 966–975. https://doi.org/10.1109/tvcg.2021.3114862

[54] Emily Wall, John Stasko, and Alex Endert. 2019. Toward a Design Space for Mitigating Cognitive Bias in Vis. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 111–115. https://doi.org/10.1109/visual.2019.8933611

[55] Monica L Wang, Alexis Gomes, Marielis Rosa, Phillipe Copeland, and Victor Jose Santana. 2023. A Systematic Review of Diversity, Equity, and Inclusion and Antiracism Training Studies: Findings and Future Directions. *Translational Behavioral Medicine* 14, 3 (Oct. 2023), 156–171. https://doi.org/10.1093/tbm/ibad061

[56] Xiaomei Wang, Ann M. Bisantz, Matthew L. Bolton, Lora Cavuoto, and Varun Chandola. 2020. Cognitive Work Analysis and Visualization Design for the Graduate Admission Decision Making Process. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (Dec. 2020), 815–819. https://doi.org/10.1177/1071181320641189

[57] Ryan Wesslen, Sashank Santhanam, Alireza Karduni, Isaac Cho, Samira Shaikh, and Wenwen Dou. 2019. Investigating Effects of Visual Anchors on Decision-Making about Misinformation. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 161–171. https://doi.org/10.1111/cgf.13679

[58] Sang Eun Woo, James M. LeBreton, Melissa G. Keith, and Louis Tay. 2023. Bias, Fairness, and Validity in Graduate-School Admissions: A Psychometric Perspective. *Perspectives on Psychological Science* 18, 1 (2023), 3–31. https://doi.org/10.1177/17456916211055374

RESEARCH MATERIAL STATEMENTS

Supplemental materials of the paper are available at https://osf.io/t6hjd/.

AUTHORSHIP

Yanan Da: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft; **Yutong Bu**: Data Curatio, Formal analysis, Writing - Original Draft, Visualization; **Yiling Li**: Data Curation, Writing - Original Draft; **Emily Wall**: Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition.

LICENSE

This work is licensed under a Creative Commons Attribution 4.0 International License.

CONFLICT OF INTEREST

The authors declare that there are no competing interests.