

BiasEye: A Bias-Aware Real-time Interactive Material Screening System for Impartial Candidate Assessment

Qianyu Liu

liuqy@shanghaitech.edu.cn

School of Information Science and
Technology, ShanghaiTech University
Shanghai, China

Qiushi Han

hanqsh@mail2.sysu.edu.cn

School of Artificial Intelligence, Sun
Yat-sen University
Zhuhai, China

Haoran Jiang

jianghr@shanghaitech.edu.cn

School of Information Science and
Technology, ShanghaiTech University
Shanghai, China

Zhenhui Peng

pengzhh29@mail.sysu.edu.cn

School of Artificial Intelligence, Sun
Yat-sen University
Zhuhai, China

Zihao Pan

panzh@shanghaitech.edu.cn

School of Information Science and
Technology, ShanghaiTech University
Shanghai, China

Quan Li*

liquan@shanghaitech.edu.cn

School of Information Science and
Technology, ShanghaiTech University,
and Shanghai Engineering Research
Center of Intelligent Vision and
Imaging, China
Shanghai, China

ABSTRACT

In the process of evaluating competencies for job or student recruitment through material screening, decision-makers can be influenced by inherent cognitive biases, such as the screening order or anchoring information, leading to inconsistent outcomes. To tackle this challenge, we conducted interviews with seven experts to understand their challenges and needs for support in the screening process. Building on their insights, we introduce *BiasEye*, a bias-aware real-time interactive material screening visualization system. *BiasEye* enhances awareness of cognitive biases by improving information accessibility and transparency. It also aids users in identifying and mitigating biases through a machine learning (ML) approach that models individual screening preferences. Findings from a mixed-design user study with 20 participants demonstrate that, compared to a baseline system lacking our bias-aware features, *BiasEye* increases participants' bias awareness and boosts their confidence in making final decisions. At last, we discuss the potential of ML and visualization in mitigating biases during human decision-making tasks.

CCS CONCEPTS

- Human-centered computing → Human computer interaction (HCI); Visualization; User studies.

KEYWORDS

bias-aware design, inconsistent decision, raise bias awareness, material screening in holistic review

*The corresponding author.

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IUI '24, March 18–21, 2024, Greenville, SC, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0508-3/24/03.

<https://doi.org/10.1145/3640543.3645166>

ACM Reference Format:

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. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA*. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3640543.3645166>

1 INTRODUCTION

The process of material screening during admissions plays a vital role in the intricate decision-making process for both college enrollment and corporate recruitment. Typically, this process involves independent reviews of various segments of the applicant's materials, resulting in a multidimensional assessment of their qualifications. Subsequently, reviewers record key points on a decision sheet for each application [56].

Application materials encompass a diverse range of documents, including personal resumes, additional certifications and letters of recommendation, among others. Given the substantial volume of applications, various automated techniques have emerged to assist in systematically and efficiently extracting and storing information. These techniques include academic exploration [25, 55] and commercial solutions such as *Daxtra*¹ and *Bello AI*². In terms of material screening, computer programs can provide a more objective and consistent assessment method based on predefined criteria [34]. They are also employed to achieve diversity in candidate selection [26]. However, these automated methods cannot comprehensively evaluate an applicant's personality and potential, nor can they fully grasp the complexities of background information. Human reviewers, on the other hand, excel at flexible adaptation [32]. Still, they are susceptible to **cognitive biases** stemming from perceptual illusions, false memories, logical fallacies and cognitive errors [32]. These biases are inherent in human perceptual and intuitive decision-making processes. While efforts can be made to

¹<http://www.daxtra.cn/>

²<https://www.belloai.com/>

identify and mitigate these biases, they cannot be entirely eliminated [32]. Furthermore, cognitive biases can be exacerbated by factors such as decision fatigue [45] and choice overload [11].

During material screening, reviewers suffer from cognitive biases stemming from several challenges. These challenges also underscore the difficulty of raising awareness about and mitigating these biases in the decision-making process. First, **the lengthy and intermittent screening process can lead to *recency bias***³ [51], as memory of earlier assessments fades over time, and decision-making criteria can be erratic (*I1*). Second, **certain attributes of applicants can trigger “halo” or “horns” effect**⁴, hindering reviewers from providing unbiased assessments of other traits [24]. For instance, during the initial assessment of academic factors, reviewers might encounter an outstanding achievement, like a perfect math grade. This initial impression can lead to an *anchoring bias* [58], where reviewers may expect equally exceptional performance in other areas then potentially lead to a biased evaluation of overall aptitude (*I2*). Third, **balancing multiple admission goals including inclusivity and selectivity is challenging** due to memory limitations and cognitive workload. Reviewers may fall prey to the *contrast bias* [50], where their judgments are influenced by the scores given to the adjacent applicants (*I3*). Lastly, forementioned challenges necessitate **inevitable revisions in the material screening process**. Reviewers often need to manually revise scores by reopening applicant pages, which can be challenging due to memory issues and *confirmation bias*⁵ [8] when revisiting applications consciously (*I4*).

Artificial Intelligence (AI) approaches, while incapable of fully replacing human decision-making in college admissions, serve as valuable tools in addressing and mitigating various cognitive biases. Previous studies in this domain can be classified into several key categories based on the life cycle of bias [15]: 1) *Prevent*. Preventative training approaches [10, 23] aim to explicitly raise awareness of bias, although they can impose a significant cognitive burden on users. Procedural interventions, on the other hand, integrate bias awareness into the decision-making workflow by enhancing information transparency [67] or providing relevant information to reviewers. 2) *Discover* and 3) *Locate*. Researchers have developed models to detect biases in real-time [22, 38, 62] and communicate these biases to users through visual elements [40, 64], based on the definitions of different cognitive biases. 4) *Mitigate*. Mitigation strategies and algorithms can be introduced based on machine learning methods [1, 22, 51] or visual approaches [13, 53, 54, 65], offering promising avenues to reduce cognitive bias. However, existing modeling methods often **target specific defined biases, neglecting the interaction between cognitive biases** [32] (research gap **RG1**). For example, anchoring bias from the earliest applications and recency bias from the recent applications may affect next decisions in the same or opposite way. Regardless of the type of bias, they can lead to inconsistent decision outcomes, as illustrated in Figure 1. In college admissions, these inconsistent screening results may conflict with the principle of *individual fairness* [22], where

³The tendency to be excessively affected by the pattern of recent data.

⁴The Halo/Horns effect is the idea that one’s perception of someone is positively/negatively influenced by his/her opinion of that person’s related traits.

⁵The tendency to favor information that supports existing beliefs while disregarding contradictory evidence.

individuals may apply different criteria at different stages of a decision task, resulting in instances with similar characteristics being treated disparately. Previous studies on fairness and diversity in college admissions [26, 34] primarily focus on the rationality of final admission outcomes, **overlooking personal inconsistent outcomes and the individual material screening process (RG2)**. Regarding visualization approaches, previous research [53, 54, 56] has demonstrated the potential of visual and interactive strategies to enhance human decision-making theoretically. Nonetheless, the **integration of AI methods and visualization strategies to address cognitive biases was infrequent**, and there was limited assessment of their combined effectiveness in practical applications (**RG3**).

To explore the factors contributing to inconsistent decision-making outcomes and the needs of reviewer for a feasible screening system, we conducted interviews with seven experienced reviewers from various academic disciplines in local universities. Based on six findings obtained from these interviews (subsection 3.2), we identified four primary challenges regarding details about four cognitive biases (introduced as **C1-C4** in subsection 3.3). In light of our literature review and identified challenges, in subsection 3.4 we devised a four-step pipeline, *PREVENTING → DISCOVERING → LOCATING → MITIGATING (RG1)*, applying to inconsistent decision-making that results from any cognitive biases, along with five essential design requirements for developing an effective system. Subsequently, we conceptualized and developed *BiasEye*, a bias-aware real-time interactive material screening visualization system. *BiasEye* served the purpose of prompting, tracking, and scrutinizing individual decision-making (**RG2**) during screening process in accordance with the four-step pipeline. The system’s backend employed *ChatGPT-4* to extract features from application materials and models individual screening preferences through a machine learning (ML) approach (**RG3**). On the frontend, *BiasEye* offered a side view that visualizes statistics for a group of applications, with each application being highlighted, as well as a summary page for retrospective decision inspection and adjustment. To assess the utility and effectiveness of *BiasEye*, we conducted a mixed-subjects user study involving 20 participants (**RG3**). The study provided strong support for the enhanced usefulness and effectiveness of *BiasEye* compared to a baseline system and any combination of baseline systems with the addition of the side view or a summary page. Notably, *BiasEye* helped participants implicitly reduce their inconsistent screening results without introducing or suggesting cognitive bias explicitly. Although the additional design elements increased cognitive load, participants reported increased confidence in their screening results’ perceived reasonability and consistency. Furthermore, the system aided participants in better establishing their evaluative criteria, resulting in more concentrated scoring for high-quality applicants within the same group. Additionally, our observations indicated that *BiasEye* facilitated participants’ understanding and explanation of the model predictions. The presence of convincing evidence played a crucial role in their final level of trust in the system’s predictions. Building upon our findings, we put forth several design implications for future developments in material screening systems. The main contributions of this study include:

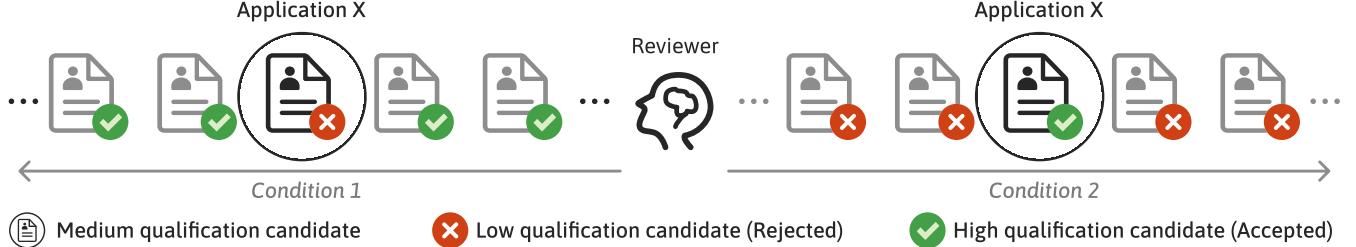


Figure 1: An illustration of the contrast bias emerges in sequential material screening tasks. In such scenarios, condition of adjacent application materials can influence a reviewer’s assessment, resulting in reviewers making inconsistent judgments about the same application X under varying conditions.

- In a formative study with seven participants, we identified challenges leading to cognitive biases in material screening and proposed a four-step pipeline to address them.
- We developed *BiasEye*, an interactive screening system that models decision preferences and mitigates bias.
- A user study with 20 participants evaluated *BiasEye*’s usability, effectiveness, and impact on behavior, workload, and confidence in screening outcomes.

2 RELATED WORK

2.1 Material Screening During the College Admission Review Process

In college admissions, the holistic review approach has been widely recognized and explored across various domains [36, 57]. A critical component of this process is material screening, which occurs after the application submission and precedes the committee meeting. Talkad Sukumar et al. [56] conducted an in-depth study on the holistic review process employed by American universities, with a specific focus on aspects related to human-computer interaction and technical support. As Sukumar et al. described, application reviewers are entrusted with a pivotal phase known as *Material Screening*, where reviewers draw upon their expertise and apply a predefined set of criteria, aligned with the university’s mission and objectives, to evaluate applications. This comprehensive evaluation encompasses a wide array of factors gleaned from the materials submitted by applicants, including a student’s high school background, family history, encountered challenges, as well as both academic and non-academic achievements, such as community service and special talents [56]. The material screening process is inherently subjective and intricate. It requires reviewers to assess applicants within the broader context of their individual backgrounds and life experiences. Rather than following a rigid, predefined protocol, reviewers rely on flexible personal heuristics, however such subjectivity would inadvertently introduce systematic errors or biases [58] such as anchoring and confirmation bias [56]. This study (subsection 3.3) will explore how four task challenges associated with four prominent biases affect the screening process and examine the tools and methodologies employed by reviewers, providing valuable insights into this crucial stage of college admissions.

2.2 Human Bias Detection and Mitigation

The college admissions screening process has low validity, limiting the ability to discern patterns and develop accurate intuitions, making experts prone to cognitive biases like anchoring

bias [12, 47, 49, 59], attraction effect [19], and confirmation bias [8]. These biases have been extensively studied and categorized in comprehensive taxonomies [20, 43, 46, 62]. Additionally, research shows that the order of presenting the same information can significantly influence decision-making [1], recent personal decisions can serve as anchors, leading to errors or inconsistencies when reviewing the same case [22]. Aligned with [22], we advocate for *individual fairness*, ensuring similar individuals are treated equitably while extends beyond addressing anchoring bias. In this study, we use “**inconsistent**” to describe situations where individual fairness is violated within the material screening process.

Detecting and mitigating cognitive bias is crucial in decision-making processes, and previous work falls into four distinct categories: **1) Preventing.** Some studies [10, 23] have focused on prevention by utilizing training approaches to raise awareness and discourage biased heuristics. However, relying solely on prior knowledge may not effectively mitigate biases and can impose cognitive burdens on users [10, 23]. Procedural interventions integrate bias avoidance into workflows without explicitly highlighting biases, such as increasing information transparency [67] and providing more relevant information to assess applicants, thereby improving the retrievability of relevant instances [56]. **2) Discovering.** Researchers have used machine learning and visual environments [43, 51] to detect human biases, some have defined and measured bias indicators [61, 62]. This category is closely associated with the next: **3) Locating.** Studies such as [64] and [40] visualized bias indicators within situational or peripheral view to pinpoint the source of bias. Echterhoff et al. [22] captured a reviewer’s anchoring state using a probabilistic model to retrospectively locate biased decisions. **4) Mitigating.** Akl et al. [1] developed strategies to reduce order-effects and enhance decision-making based on probability models. Visual methods, such as design spaces [65] and simple visual representations [13], have been proposed to mitigate cognitive bias. Researches [53, 54] have demonstrated that implementing visualizations in the review process can automatically address cognitive biases, alleviating user concerns.

In this study, drawing inspiration from prior research, we have integrated a four-step pipeline into our material screening system. First, we present supplementary information and statistics related to applications to **prevent** cognitive bias. Next, employing machine learning techniques, we create dynamic models of real-time individual decision preferences based on a user’s historical choices. Through our visualization design, users can **discover** and **locate** any inconsistencies in their decisions, ultimately helping them **mitigate** these inconsistencies conveniently.

2.3 AI-Enhanced Approaches for Material Screening and Holistic Review Support

Material screening serves as the crucial initial step in assessing a candidate's qualifications. To enhance efficiency and fairness, various methods have been developed to optimize procedures such as the holistic admission process [31] and information ordering [2], or automate particular tasks such as resume screening [48], assessment [34], and information extraction [30, 35]. Natural Language Processing (NLP) [28] also has been used to detect and correct resume errors [41] and conduct rating classification while reducing human bias [3].

Although automated screening can mitigate human bias, concerns about potential discrimination stemming from biased data or algorithms, including racial discrimination, have been raised [16–18, 42]. Initiatives like FairCVtest [44], MANI-Rank [9], and Gilbert et al.'s human-centered AI tool [26] aim to address these issues. However, it's important to note that while these methods automate parts of application material processing, they may not fully capture human review patterns or contextual nuances, limiting their use in holistic admissions reviews. Our approach uses machine learning as a supportive tool for human decision-making, adapting to individual reviewer preferences to personalize bias mitigation while retaining the final decision in the hands of the human reviewer.

Several software platforms, such as *Slate*, *Kira Talent*, and *Submittable* [52], support holistic review processes. The American Association of Medical Colleges (AAMC) also offers tools and principles for holistic review [4]. Additionally, the College Board and Education Counsel jointly published a guide [14] that includes a diversity metrics dashboard. Metoyer et al. [39] explored group decision-making and integrated visual storytelling support into collaborative review for transparency and rigor. While these efforts focus on addressing cognitive bias and human decision-making in holistic admissions, our study centers on the material screening process prior to committee meetings. We aim to enhance bias awareness and promote self-reflection through machine learning and visualization in digital applications, building models of reviewers' personal decision preferences.

3 FORMATIVE STUDY

This study aims to help reviewers deal with inconsistent review decisions caused by cognitive biases. To achieve this, we conduct a formative study to understand reviewers' current practices and needs. These insights will inform the design requirements for a system tailored to this context.

3.1 Participants and Procedure

To comprehensively understand the current state-of-the-art material screening process, challenges faced by reviewers, and their expectations for screening systems, we conducted semi-structured interviews with seven experienced reviewers. The participants, with a mean age of 32.6 years (standard deviation 13.1), included four males and three females, offering diverse perspectives. Our objectives were twofold: to explore current practices and challenges in material screening and identify strategies to mitigate cognitive biases while enhancing efficiency and satisfaction. Interviewees represented various roles, including admissions officers, material

reviewers, and interviewers, spanning academic and professional backgrounds in fields like Computer Science, Industrial Design, Entrepreneurial Finance, and FinTech. We developed the interview script through informal discussions with admission officers and reviewers. As outlined in Table 1, participants discussed their screening procedures, experiences, and shared views on four cognitive biases: *anchoring bias*, *recency bias*, *contrast bias*, and *confirmation bias*, along with coping strategies and specific requirements within each scenario.

We used Braun and Clarke's six-phase thematic analysis framework [27] to analyze interview data. The analysis involved two researchers proficient in qualitative research methods. One researcher performed the initial coding of the data, while the other meticulously reviewed the codes to ensure accuracy and completeness. Through iterative discussions, two authors reached a consensus on the summarizing statements at first, resolving potential ambiguities or conflicts. Next, they collaboratively identified six screening findings together, subsequently giving rise to four key challenge themes discussed in subsection 3.2 and subsection 3.3, respectively. These insights informed the derivation of five design requirements, forming the foundation for a four-step strategy explained in subsection 3.4.

3.2 Findings about the Current Material Screening Process

This section presents six key findings from our interviews about the current material screening process, comparing them with findings in [56].

Finding 1: Multiple rounds of material screening. Material screening has become more complex and time-consuming, with universities adopting a multi-round approach (E1, E2, E5), differing from the simplified approach in [56] where one reviewer was assigned per applicant. Moreover, reviewers encompass a spectrum of experience levels, ranging from senior assistant students acting as junior reviewers to professors serving as expert reviewers in each evaluation cycle. This approach achieves a dual objective of maintaining selectivity and inclusivity simultaneously. As E1 noted, “*A significant number of applications exist, and junior reviewers should screen out the underperforming ones, thus allowing expert reviewers to focus on the more competitive submissions.*” Applicants undergo multiple reviews leading to interviews and committee meetings to finalize admissions.

Finding 2: Multiple reviewers in each round. To mitigate the impact of personal preferences, each applicant is assigned to reviewers from various departments, and their scores are averaged to determine the effective score (E2, E4, E6, E7). According to E6, “*Reviewers possess their own preferences, and enabling reviewers with diverse backgrounds to assess the same applicants aligns with the objective of achieving a more diversified admissions process.*” Similar to [56], reviewers primarily handle applications from their respective or familiar regions but not exclusively so.

Finding 3: Diverse admission expectations. As noted in [56], reviewers are tasked with balancing diverse and inclusive admission goals with the school's mission. Moreover, fair assessment is ensured by considering the average scores from at least three reviewers per round. Assuming a normal quality distribution, admission

Category	Question
Demographic	How many times have you participated in material screening or interviews for college admissions?
	What is the overall flow of college admissions?
	What policies/criteria has the admissions committee formulated?
Procedures	Who is qualified to be a reviewer? (How the Admissions Committee recruited the reviewers?)
	How the applications were distributed to reviewers?
	How to evaluate an application comprehensively and decide admission results?
Prompt	What are the functions of current screening system?
	Have you ever overestimated or underestimated applications?
	How did you handle this problem and avoid the similar situation?
Scenarios	(1) Do you think that certain aspect of the applications will affect your assessment of their other aspects? [anchoring bias]
	(2) Do you think the review process is affected by time and memory? [recency bias]
	(3) Do you think the sequence order of applications may affect your assessment? [contrast bias]
	(4) Did you objectively assess the shortcomings of applications when you have had a favorable impression of them? [confirmation bias]
Expectation	What functions do you want to add or improve to the current screening/interview system?

Table 1: Interview with expertise reviewers.

committee manages a large volume of applications by randomly distributing and sequencing them. As noted by E3, E4, and E7, reviewers are instructed to target a suggested mean score, mitigating aggregation errors.

Finding 4: Flexibility in reviewer work schedules. Reviewers are provided with one to two weeks to autonomously accomplish their screening assignments, usually during breaks in their regular work and study schedules, as outlined in [56]. Reviewers have the flexibility to either assess a few applications daily during their spare moments or allocate a dedicated continuous time block to evaluate all applications (E1-7).

Finding 5: Aggregating multi-dimensional assessments. Candidate assessment involves considering various dimensions such as educational background, academic and non-academic activities, and letters of recommendation [56]. Universities assign weights to each dimension for an overall score, rather than a single cumulative score. Furthermore, the admission office provides a list of competitions or awards for seamless integration into the scoring system, as E4 mentioned, “*This process has been automated recently as part of the system iteration.*”

Finding 6: Outdated material screening system. As discussed in [56], existing material screening systems are predominantly representational and lack interactivity. Reviewers navigate a list of applications, each with bundled PDF materials and a digital decision sheet for scoring and commenting. The application list shows screening progress, scores, and a submission button. While these systems provide basic functionalities like electronic storage and accessibility, they lack advanced features.

3.3 Challenges in Material Screening Process

In this section, we will explore each challenge themes (**C1-C4**) by examining the fundamental characteristics (*Finding 1~6*) of the material screening process, helping us identify potential cognitive biases in this phase.

C1: Balancing workload and fairness in college admissions screening. Despite the need for a holistic approach in college admissions (*Finding 3*), the high volume of applications often restricts

the time and energy reviewers can dedicate to each student, hindering thorough exploration and deliberation (E1, E3, E4, E6). The automated awards-to-score approach in *Finding 5* may reduce some workload, but not all awards are listed, and subjective judgment based on experiential knowledge remains necessary. As E7 stated, aligning average students with score distribution requirements in *Finding 3* is challenging, “*How do you come up with the boundary for those average students? It's a bit tricky, and honestly, I didn't know it right from the beginning.*” The *contrast bias*[50] is particularly evident with intermediate qualifications, an outstanding applicant can overshadow others, and a series of subpar materials may lead to higher scores for an average applicant [22].

C2: The screening procedure can be quite time-consuming and frequently intermittent. As highlighted in *Finding 4*, the screening task is susceptible to interruptions and places a substantial memory burden on reviewers due to their constrained time (E4, E6) and the fragmented nature of their personal schedules (E1, E2, E3). Moreover, the influence of *recency bias* [51] prompts reviewers to base their decisions on applicants they’ve recently assessed (E5), resulting in a fluctuating personal evaluation criterion.

C3: Reviewers might be susceptible to the allure of the halo effect⁶. As *Finding 5* and [56] suggest, an applicant’s academic performance can anchor assessments of other dimensions (E1, E3, E4), validating the *anchoring bias* [58]. Furthermore this anchoring effect can positively or negatively manifest in various aspects. For instance, E3 remarked, “*This student possesses extensive experience and stands out among applicants. Excellent! I'm inclined to award extra points in every dimension.*” Conversely, E4 expressed doubts, stating, “*Did his parents ghostwrite this self-introduction? Some phrases appear to be readily available online, suggesting a lack of sincerity, which raises concerns about the other achievements.*” Anchoring bias can subtly influence reviewers using a heuristic approach to decision-making, resulting in unintentional inconsistent outcomes.

⁶The Halo effect means one’s perception of someone is positively influenced by his/her opinion of that person’s related traits.

C4: Reviewers struggle with inconvenient systems and lack guidance when making score adjustments. Not discussed in [56], initial lower scores [5] resulted from reviewers' caution due to incomplete understanding. As E7 expressed, “*I acknowledge that this may seem unfair to students in the front. I'll proactively make adjustments, although I can't guarantee them.*” As screening progressed, decision fatigue led to declining decision quality and preference for expedient or mean-score heuristics (E4). Inconsistent outcomes (C1-C3) necessitated revisions, with reviewers revisiting decisions repeatedly and verifying before final submissions. E1 emphasized iterative score revisions to ensure fairness, stating, “*It's essential to make adjustments, especially when more competitive performances are noticed further back. I need to lower those high scores at the front.*” However, this was exhausting due to unreliable memory and the outdated system (*Finding 6*). E6 suggested making retrospective assessment more intuitive and optimizing interaction beyond the current “click to display” method.

3.4 A Four-Step Pipeline and Design Goals

Drawing from relevant research and interview insights, we present a four-step pipeline to address four challenges and providing a foundation for system design $\mathcal{D}1\sim\mathcal{D}5$.

Step 1: Preventing. Humans may not consistently excel at repetitive tasks [21], so enhancing screening quality, especially addressing C1, involves reducing the reviewer's workload. **This means focusing on automating routine tasks, allowing reviewers to devote their attention to more complex subjective assessments ($\mathcal{D}1$).** To tackle C1, $\mathcal{D}1$ includes preprocessing and gathering necessary information for screening applications, thus enhancing the retrieval of instances related to the availability heuristic [56]. Furthermore, automating repetitive judgments and actions through intuitive representation and simplified interaction is a practical strategy for C4.

Step 2: Discovering. Screening procedure and human cognitive process constraints can lead to biased decisions (C1-C3) [32]. However, participants were either unaware of or underestimated bias impact in their assessments (E4, E7). With subjective criteria involving intangible, shifting factors, an objective approach is needed to help recognize and rectify irrational behavior. Our system aims to **facilitate reviewers' understanding and management of the screening process ($\mathcal{D}2$)**, as well as **explicitly reveal screening preferences to uncover potential inconsistencies ($\mathcal{D}3$)**.

Step 3: Locating. While the initial *discovery* step offers an overview of the screening process, in-depth analysis is crucial to implement targeted strategies addressing inconsistencies from C1-C3. $\mathcal{D}4$ involves **examining bias tendencies and evaluating specific bias instances**. Our system should provide transparent, comprehensive information for multifaceted material comparisons. Through interactive visualization, reviewers can identify inconsistent outcomes and make informed judgments, promoting fairness and objectivity in screening.

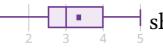
Step 4: Mitigating. The final key step in bias mitigation is score modification. The current system, mainly representational, lacks interactivity, burdening reviewers physically and cognitively when adjusting decisions (C4). Additionally, comparing numerous similar candidates for reasoned scores is challenging (E7). To address C4, our system should **enable quick adjustments and provide**

reasonable score recommendations ($\mathcal{D}5$) to ease reviewers' bias concerns. Meanwhile, interactive visualization is a promising way to enhance assessment efficiency and effectiveness.

4 BIASEYE

In line with design goals $\mathcal{D}1\sim\mathcal{D}5$, we present *BiasEye*, a real-time interactive system aimed at assisting reviewers in preventing, discovering, locating, and mitigating inconsistent decision-making. Implemented using Flask and Vue.js frameworks, it leverages Element-plus components⁷ and D3.js⁸ [6] for visualization. *BiasEye* consists of three pages: 1) *Student List*, displaying assigned applications and screening progress; 2) *Assessing*, showing extracted information and original PDF materials for application assessment; 3) *Summary*, offering retrospective bias-aware score inspection and revision through the *Screening Sheet*, *Comparison* view, and *Ex-situ Table*. All three pages share a *Statistical* view accessible via header navigation (Figure 2-a). Potential usage methods and scenarios are explored in subsubsection 6.2.1 to address inconsistent outcomes.

4.1 Screening Sheet

Aligning with the admission committee's criteria, the *Screening Sheet* includes a *Basic Information* section and several screening sections, each with a unique color. In addition to the score⁹  3 and the comment component involved in the original decision sheet mentioned in *Finding 6*, each section also showcases structured entries extracted from resumes and included a box plot  showing statistical data for the assigned scores.

As depicted in Figure 3, we first convert PDF files into TXT format and filter out resumes with incomplete or inaccurate information, ensuring the quality of our analysis. To extract information, we explored models and tools like *CNN-BiLSTM-CRF* [35] and *pyresparser*¹⁰, but these had suboptimal performance due to diverse resume formats and limited training samples. Consequently, we fine-tuned *ChatGPT-4*¹¹, implementing error correction codes and human verification for precision and consistency. Despite limitations like incomplete extraction of low-probability information with limited training data, this tool was effective in resume information extraction. Finally, as depicted in Table 2, all raw text was extracted and structured into JSON format, encompassing five sections: *Basic Information*, *Educational Background*, *Competition*, *Honors*, and *Extra Activity*. Letters of Recommendation (LoR) and Personal Statements (PS) are not displayed in this sheet, but users can access these original files directly from the *Assessing* page.

Considering $\mathcal{D}1$ in the *PREVENTING* step and $\mathcal{D}4$ in the *LOCATING* step, we incorporate the *Screening Sheet* with user-friendly interactivity into the *Summary* page. This allows reviewers quick access to concise information about the selected applicant. While some details may be missing compared to the original PDF, the sheet provides ample cues. If more information is needed, reviewers can switch to the *Assessing* page to examine the PDF.

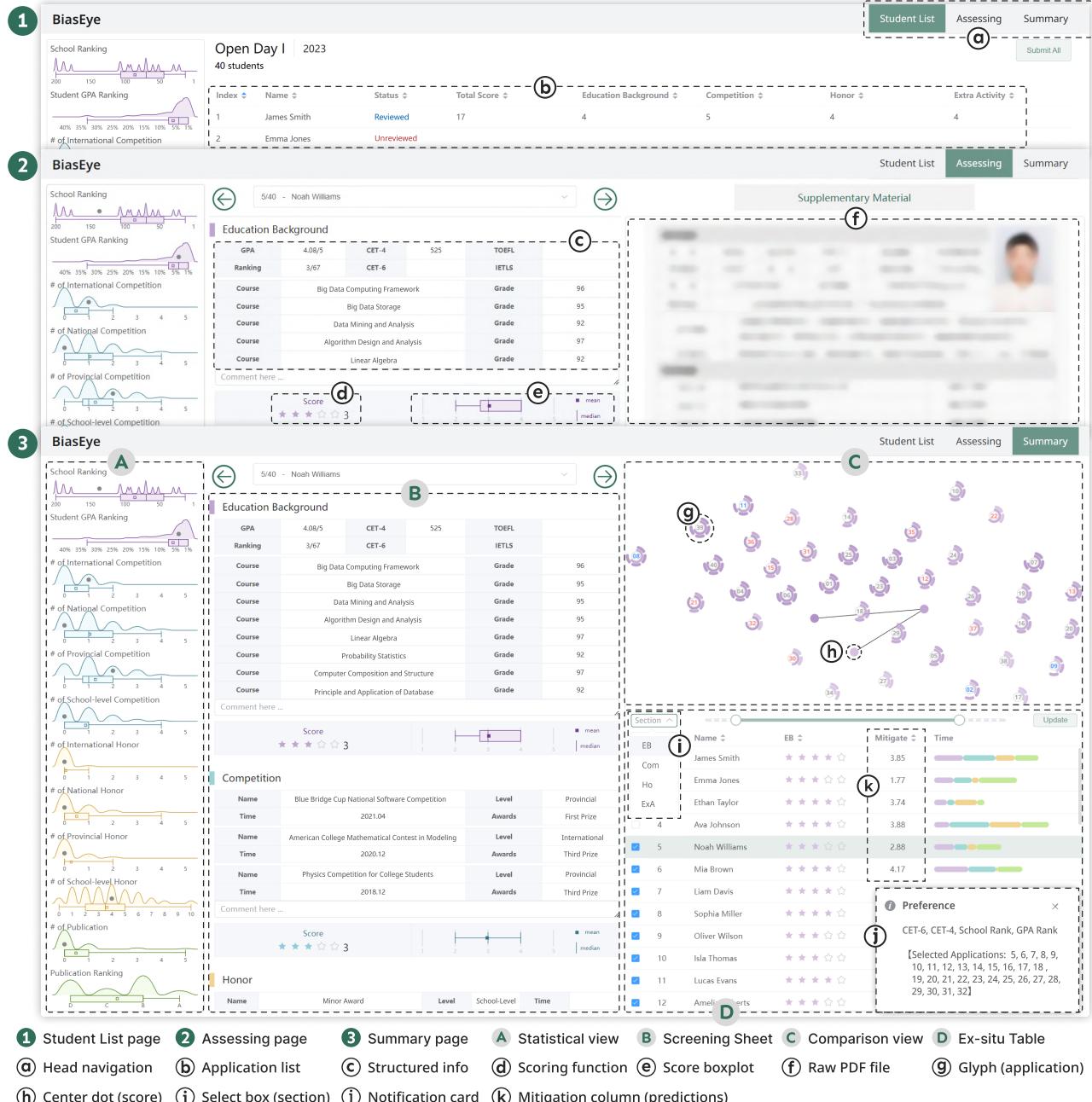
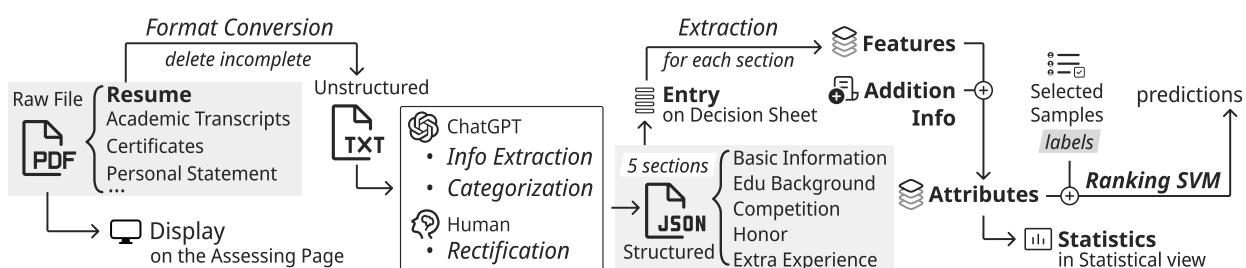
⁷<https://element-plus.org/>

⁸<https://d3js.org/>

⁹While each section employs a consistent score range of 1-5 points, the system assigns different weights to each section when computing the total score, as per the admission committee's guidelines.

¹⁰<https://github.com/OmkarPathak/pyresparser>

¹¹<https://chat.openai.com>

Figure 2: The front-end design of *BiasEye*.Figure 3: The overview of data processing and backend model pipeline of *BiasEye*.

Section	Entries
Basic Information	Name, Gender, Hometown, School, Major, Skill
Education Background	GPA, Student Rank, CET-4, CET-6, TOFEL, IELTS, Course Name, Course Grade
Competition(*)	Name, Time, Level, Award
Honor(*)	Name, Time, Level
Extra Activity(*)	Project: Name, Time, Role, Description Research Paper: Title, Author (order), Publication, Level, Summary Other Experience: Name, Time

Table 2: Structured information entries from resumes in JSON formats, the extra activity section are divided into three sub-categories. *: The corresponding entries represent the content of each record in that section.

Section	Attributes	
Education Background	CET-4, CET-6, TOFEL, IELTS	
Competition (#)	School Award, Provincial Award, National Award, International Award, Mathematics Competition, English Competition, Computer Competition, Chemistry Competition, Electronics Competition, Mechanical Competition, Physics Competition, Biology Competition, Innovation and Entrepreneurship Competition, Other Competition	School Rank
Honor (#)	School Honor, Provincial Honor, National Honor, International Honor, Scholarship, Excellent Student, Outstanding Student, Outstanding Graduate, Student Officer, Volunteer, Social Practice, Skill Certificate	Student Rank*
Extra Activity (#)	A-tier Publication, B-tier Publication, C-tier Publication, D-tier Publication, Projects, Project Manager, Project Participant	

Table 3: Attributes for each screening section, four sections share the attributes of School Rank and Student Rank. #: The corresponding attributes represent the quantitative outcomes following aggregation. *: Indicates that the attribute has been normalized.

4.2 Statistical View

In response to $\mathcal{D}1$, we design the *Statistical* view (Figure 2-A) on *BiasEye*'s left side. This view presents global statistics for the current application group, visualizing 12 key indicators. These include school and normalized student GPA rankings, competition and honor count at various levels, and publication counts with corresponding conference/journal levels. Each indicator (Figure 4) uses box plots to convey central tendencies and data dispersion, density plots to offer detailed distributional insights, and scatter dots to depict the cases of the currently selected students, offering an overarching perspective that aids *PREVENTING* recency and contrast bias.

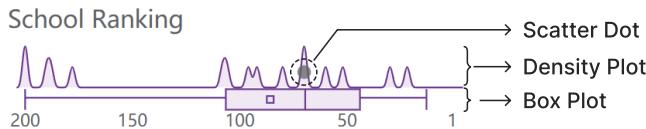


Figure 4: A visualized indicator of the *Statistical* view.

For each section, we defined a set of significant attributes denoted as $A = a_1, a_2, \dots, a_M$ based on feedback obtained during the formative study. These attributes serve as straightforward proxies for human decision-making preferences, as outlined in Table 3. On one hand, most of these attributes are derived from features extracted directly from entries within the JSON file. Numerical and quantitative features, such as the count of different competition levels, can be readily obtained from the respective entries. Some features necessitate a text classification step before quantitative calculations, such as determining whether an applicant served as a manager or participant in a project. To facilitate this, we presented

input and output samples to *ChatGPT* and guided it through the classification process, providing explanations along the way. This approach aimed to encourage *ChatGPT* to engage in a more deliberate thought process. On the other hand, two attributes, namely school ranking and publication ranking, were derived from additional information. This addresses $\mathcal{D}1$ and aims to streamline the information search process, ultimately reducing the reviewer's workload. The school ranking is assigned a label from 1 to 200¹², while the conference/journal level is categorized from A to D¹³, where 'D' signifies 'unknown'.

4.3 Ex-situ Table



Figure 5: A stacked time bar *Ex-situ* Table.

As an extra enhanced version of the *Student List* table, *Ex-situ Table* incorporates additional visualizations and interactive features to address $\mathcal{D}2$, $\mathcal{D}4$ and $\mathcal{D}5$. It provides an overview and facilitates score modifications, displaying application ID, applicant name, and section duration, which are calculated from time difference between two consecutive scoring events. Hovering over a stacked bar (Figure 5) in 'Time' column reveals specific time values. Clicking a row in the table updates the *Screening Sheet* and highlights the corresponding application glyph in the *Comparison* view. The table

¹²University rankings: <https://research.com/university-rankings/best-global-universities>

¹³Conference/journal level: <https://research.com/>

dynamically displays the corresponding section column based on the selection in **Section** , enabling direct score modification for $\mathcal{D}5$. Additionally, the *Ex-situ Table* offers an interface that employs a machine learning method, specifically, *Ranking SVM*, to help users *DISCOVER* ($\mathcal{D}2$) inconsistent decision outcomes for each screening section. Through the use of a slider  and checkboxes , users can select a specific number of assessed applications as trusted training samples. Clicking the **Update** button activates the *Ranking SVM* for analysis.

Inspired by *Podium* [63], we employed *Ranking SVM* [33] to automatically infer attribute weights from user-assigned application scores. This approach serves two purposes: Firstly, it helps reviewers examine their individual screening priorities and preferences, providing insight into how personal biases and emphases may affect their assessments. Secondly, *Ranking SVM* forecasts future review tendencies using past records, indicating potential biases. Its low computational cost enables real-time monitoring, allowing reviewers to make timely adjustments and evaluate the appropriateness of modifications.

Derive constraints. *Ranking SVM* optimizes a loss function involving pairwise constraints based on the Support Vector Machines (SVM) framework. We constructed a training set for the *Ranking SVM* model using a subset of $k (> 6)$ user-selected assessed applications, each assigned a score represented as S . We form pairs of data points (d_i, d_j) with a label l . If $s(d_i) < s(d_j)$, we set $l = 1$; otherwise, $l = -1$. For all pairs $i, j \in 1, \dots, k$ where $i \neq j$, we generated constraint tuples based on this criterion and treat all constraints as soft constraints.

Calculate the ranking and transfer to score. After training, we obtained a weight vector for the attributes to rank all the data items. We computed individual dot products of the weight vector (w) with each data item (d_i), resulting in an intermediate variable denoted as $v(d_i) = w \cdot d_i = \sum_{m=1}^M w_m \cdot a_m$, where a_m represents the attribute value in the selected section. Subsequently, we mapped the values of v to the interval $[min(S) - 0.5, max(S) + 0.5]$, preserving two decimal places to enhance transparency and facilitate explanation. This mapping yielded the prediction score S' , with the condition that $s'_i = 0$ if $s_i = 0$.

The prediction score is displayed in the ‘Mitigate’ column, and a notification appears in the bottom right of the page, listing the top $k (= 10)$ significant model attributes and training application IDs. Users can *LOCATE* ($\mathcal{D}4$) inconsistent decision outcomes by comprehensively comparing these predictions with their scores and cross-referencing this information with the significant attributes and original data.

4.4 Comparison View



Figure 6: Design of glyphs in *Comparison* view.

To address $\mathcal{D}1$, we developed a visual glyph (Figure 6) for comparing human scores and model predictions. Each glyph corresponds to an applicant, with the number denoting the application

ID, the outer ring encoding the human score and the inner ring encoding the prediction. A linear color scheme is used for both rings, facilitating the rapid identification of applications with varying scores. The ID color indicates whether the human score is *higher/lower* than or *close to* the prediction, highlighting inconsistencies in human scores and their direction. To *LOCATE* ($\mathcal{D}4$) anomalies among similar applications, glyph position is determined using the *t-SNE* [60] method based on the attributes of selected section, ensuring similar applications are closer together. Solid dots represent high-dimensional centers of all applicants who received the same human score, follow the same color scheme as the glyph rings and are connected from lowest to highest scores ($\mathcal{D}2$). Furthermore, to provide visual aid for $\mathcal{D}4$, hovering over a glyph or center highlights applicants with the same human score.

5 USER STUDY

Obtaining institutional IRB approval, we conducted a user study involving 20 participants with mixed backgrounds. The primary aim of this study was to assess the effectiveness of our bias-aware design. To achieve this, we aimed to address three key research questions ($RQ1 - RQ3$) through our evaluation process.

- $RQ1$: How are the usability and effectiveness of the bias-aware system in material screening?
- $RQ2$: How will participants interact with and be affected by the bias-aware system in material screening?
- $RQ3$: How will participants trust and collaborate with the ML method?

5.1 Experiment Setup

5.1.1 Dataset. We obtained IRB approval for data collection and used a dataset from a local university’s information science master’s program. Graduate admission, similar to college admission, emphasizes merit and alignment with the institution’s mission. We randomly selected two groups of 40 complete applications (excluding incomplete ones) for a preliminary trial and a formal experiment. Each application included a resume, academic transcripts, a personal statement (PS), and up to two letters of recommendation (LoR), all in PDF format. To ensure the experiment’s completion within 1.5 hours, we retained only the resume, transcripts, and certificates. Notably, these materials were from past admission interviews, and we had only raw PDF files, making it impossible to verify results with reliable ground truth. Additionally, we anonymized identifiable details like names and photos.

5.1.2 Baseline System and Control Conditions. We adopted a two-pronged approach to assess the effectiveness of our system. First, we used a **between-subject design to evaluate the Statistical view**, dividing participants into two groups randomly: Group A used the baseline system, and Group B used *BiasEye*. Both systems consisted of three pages (Figure 2-   , but the baseline system lacked the *Statistical* view and publication level in the *Screening Sheet* (Figure 13). Both systems were hosted on a web server, accessible to participants via public links. Second, we used a **within-subject design to evaluate the Summary page** in two stages. In stage I, participants could only use the *Student List* and *Assessing* pages. In stage II, participants could further adjust their decisions using the entire system.

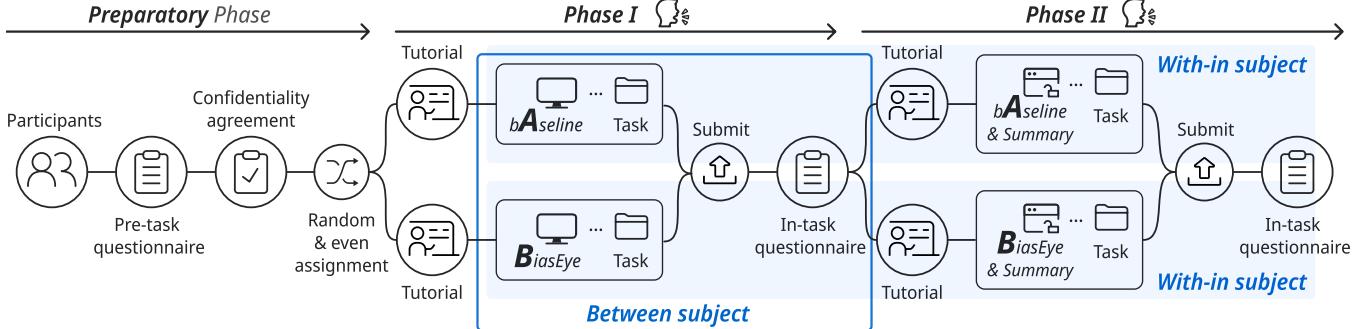


Figure 7: Procedure of user study.

ID	Gender/Age	Degree	Experienced	Condition Group	ID	Gender/Age	Degree	Experienced	Condition Group
P1	Male / 21	Bachelor	N	A	P2	Male / 23	master	Y	B
P3	Male / 24	master	N	A	P4	Male / 24	master	N	B
P5	Female / 23	master	Y	A	P6	Female / 24	master	Y	B
P7	Male / 23	master	Y	A	P8	Male / 23	master	Y	B
P9	Male / 23	master	Y	A	P10	Male / 21	master	N	B
P11	Male / 21	Bachelor	Y	A	P12	Male / 25	Ph.D	Y	B
P13	Female / 26	Ph.D	N	A	P14	Female / 24	Bachelor	N	B
P15	Male / 22	master	Y	A	P16	Male / 21	Bachelor	Y	B
P17	Male / 21	Bachelor	N	A	P18	Female / 23	master	N	B
P19	Female / 22	master	N	A	P20	Male / 21	Bachelor	N	B

Table 4: Demographic information of participants. Experienced means one has prior involvement in relevant screening assistance scenarios encompassing over 20 applications. Group A uses Baseline system, group B uses BiasEye system.

5.2 Participants

We recruited 20 participants (P1 to P20): 14 males and 6 females. Among them, 6 held bachelor's degrees, 12 held master's degrees, and 2 held Ph.Ds. Participants were evenly divided into the experiment (B) and control (A) groups based on demographics (Table 4). Before the formal experiment, all participants signed a confidentiality agreement, became familiar with the training program and department's mission. Special attention was given to those without prior relevant experience ($n > 20$) to ensure they understood the screening expectations. Their participation was incentivized by performance-based compensation.

5.3 Task and Procedure

5.3.1 Task. We simulated a real-world material screening scenario for user study. Participants were instructed to act as students in a Human-Computer Interaction (HCI) laboratory, tasked with preliminarily review 40 admission applications due to the time constraints of their professor. The participants' responsibility was to consider multiple factors like personal backgrounds, experiences, abilities, and the lab's requirements. Their anonymous screening outcomes would be combined with others to determine final screening results. To fulfill this task, participants had to: 1) assign scores to each application in four sections: *Education Background (EB)*, *Competition (Com)*, *Honor (Ho)* and *Extra Activity (ExA)*, which were chosen based on the actual department criteria. 2) They were prohibited from discussion and communication, and 3) were not required to consider score weighting within each section. 4) They were encouraged but not forced to aim for an average score of 3 in each section. Additionally, online references¹⁴ including school rankings, conference and journal rankings, and a formal document listing the level of college student competitions were provided for assistance.

5.3.2 Procedure. Figure 7 outlines our mixed-subject experiment. Before the study, participants signed confidentiality and completed a pre-task questionnaire collecting demographics. We introduced the experimental task and its objectives in a comprehensive manner, **rather than explicitly disclosing the focus on cognitive bias**, we emphasized the core principle of *individual fairness* and underscoring the gravity of inconsistent outcomes. This approach ensured that participants remained unaware of the precise nature of our study.

Next, we introduced the system corresponding to their belonging condition in stage I and provided a set of toy trial materials for familiarization. During Phase I, participants were allotted 50–70 minutes to complete the task as consistently as possible, then submitted their results and filled out an in-task questionnaire. The main goal of Phase I is to assess how the introduction of the *Statistical* view impacts the consistency of participants in decision-making. To address potential residual effects between the two experiments and reduce response bias within the two conditions, we adopted a between-subject design approach.

Moving to Phase II, we introduced the *Summary* page and the *Ranking SVM* model, which learns participants' screening preferences and predicts scores. To directly compare the change in decision-making before and after model intervention, we utilized a within-subject design independently for both groups. Simultaneously, both groups maintained a between-subject design that included the *Statistical* view as a variable. Participants were given 20 minutes to revise their outcomes with the assistance of *Summary* page. Subsequently, they submitted again and completed a post-task questionnaire.

Two of the authors acted as experimenters to ensure smooth progress and provided assistance as needed. The study spanned approximately two hours, with participants receiving USD 12 compensation on average.

¹⁴<https://research.com/>

5.4 Data Collection

We conducted a general quality check for each participants by examining the usage time of Phase I, which started when they began the task and ended at their first outcome submission. One submission from group B (P20) was rejected due to a extraordinarily short duration (30 minutes) for Phase I. Besides, one scoring log files form group A (P1) were irreversibly corrupted, we excluded his log files and questionnaires from quantitative analysis but kept video for qualitative analysis. We ended up with 18 valid responses, 9 per group. All data will be used solely for experimental outcome analysis and won't be shared or disclosed non-anonymously.

5.5 Measurement

For both the in-task and post-task questionnaire, we utilized a 7-point Likert scale (1: Not at all/Strongly disagree, 7: Very much/Strongly agree, and a 10-point scale for workload-related questions) to collect participants' feedback on the respective systems and their attitudes toward their own results in different phases of the study. First, in line with the **System Usability Scale** (SUS) [7], we crafted questions primarily including: 1) Ease of use; 2) Ease of learn; 3) System satisfaction; and 4) Likelihood of future use; Second, in terms of **Self-Evaluation**, we designed questions mainly including: 1) Consistence criteria; 2) Degree of distinction; 3) Fewer revisions; and 4) Perceived efficiency promotion. Third, drawing from the NASA-TLX survey [29], we posed questions about **Workload Assessment**, including: 1) Psychological workload; 2) Physical workload; 3) Time workload; and 4) Level of frustration. Fourth, as for **System Design**, we tailored questions concerning the *Statistical* view for group B participants in the in-task questionnaire and regarding the *Ex-situ Table* and *Comparison* view for both groups in the post-task questionnaire, including: 1) Intuitive visualization; 2) Convenience of interaction; and 3) Overall helpfulness. Additionally, we also included optional subjective questions for qualitative insights. Participants were instructed to “think aloud” throughout while their screens and audio were recorded. The system documented the section name and score for every modification in scoring logs during both phases for later quantitative analysis in section 6.

6 RESULTS AND ANALYSIS

This section organizes quantitative and qualitative results for research questions RQ1 to RQ3. Our **quantitative analysis**, besides descriptive statistics, employed the Mann-Whitney U test [37] to investigate differences between groups using different systems, and the Wilcoxon signed-rank tests [66] to evaluate disparities between groups of participants using the same system. For our **qualitative analysis**, one author transcribed participants' screen recordings, capturing system usage and reactions to potential inconsistent decisions. Two authors then coded these transcriptions using thematic analysis [27], with specific examples included in this paper.

6.1 RQ1. How are the usability and effectiveness of the bias-aware system in material screening?

As shown in Figure 8, the questionnaire presents participant ratings of system usability at various stages and with different systems. When comparing the Phase 1 data for both systems, we observed



Figure 8: The usability of usefulness of the system. Error bars indicate standard errors. (ns: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$).

that the *BiasEye* system did not lead to a significant increase in ‘ease of use’ or ‘ease of learning’. However, it did demonstrate a substantial increase in ‘satisfaction’ ($U = 3.5, p < 0.01$) and ‘future use’ ($U = 1.5, P < 0.001$).

Conducting a comparative analysis of data within the same system at different phases, we noticed that the introduction of the Summary Page had a significant impact. Specifically, it led to a decrease in ‘ease of learning’ for both the Baseline and *BiasEye* systems ($T = 0.0, p < 0.05$ in Baseline, $T = 0.0, p < 0.05$ in *BiasEye*). Furthermore, it significantly enhanced ‘satisfaction’ ($T = 0.0, p < 0.05$) in the case of the Baseline system. However, there were no significant changes in terms of ‘ease of use’ and ‘future use’ for both systems.

Moving forward, we proceed to evaluate the efficacy of the *BiasEye* system by delving into the data collected from participants as they engaged in the real scoring process. Our analysis has unveiled the following two key findings.

Finding 7: The Statistical view and additional information facilitates participants in raising awareness of bias in the process and proactively reducing inconsistencies in decision-making. Our findings stem from an examination of participants' interactions with the system, focusing on instances where they adjusted their initially assigned scores. The statistical analysis of score revisions during Phase I and Phase II is presented in Figure 9(a) and Figure 9(c), respectively.

In Figure 9(a), it becomes apparent that participants using the *BiasEye* system displayed significantly higher average frequencies of score revisions for the *EB* ($U = 522, p < 0.01$), *Ho* ($U = 539, p < 0.01$), and *Sum* ($U = 389, p < 0.001$) categories compared to those using the Baseline system. As there was no machine learning intervention in Phase I, participants adjusted their decision outcomes relying on personal judgment. The transcripts indicate that participants recognized the inconsistency in their initial decisions, and their perception of this inconsistency became more pronounced and less ambiguous. Participants demonstrated the ability to discern candidates with varying qualifications more swiftly and accurately.

To reinforce this observation, we plotted a scatter plot in Figure 9(c) depicting the average number of score changes against the applicant sequence, fitting a linear function. As depicted, as the number of students being scored increased, both groups experienced a decline in the frequency of revisions. This observation aligns with the expectation that participants' evaluation criteria improve and stabilize over time. Notably, the fitted line for *BiasEye*

users is always higher than the baseline, suggesting that the proposed system increased the number of revisions generally, rather than being influenced by outliers.

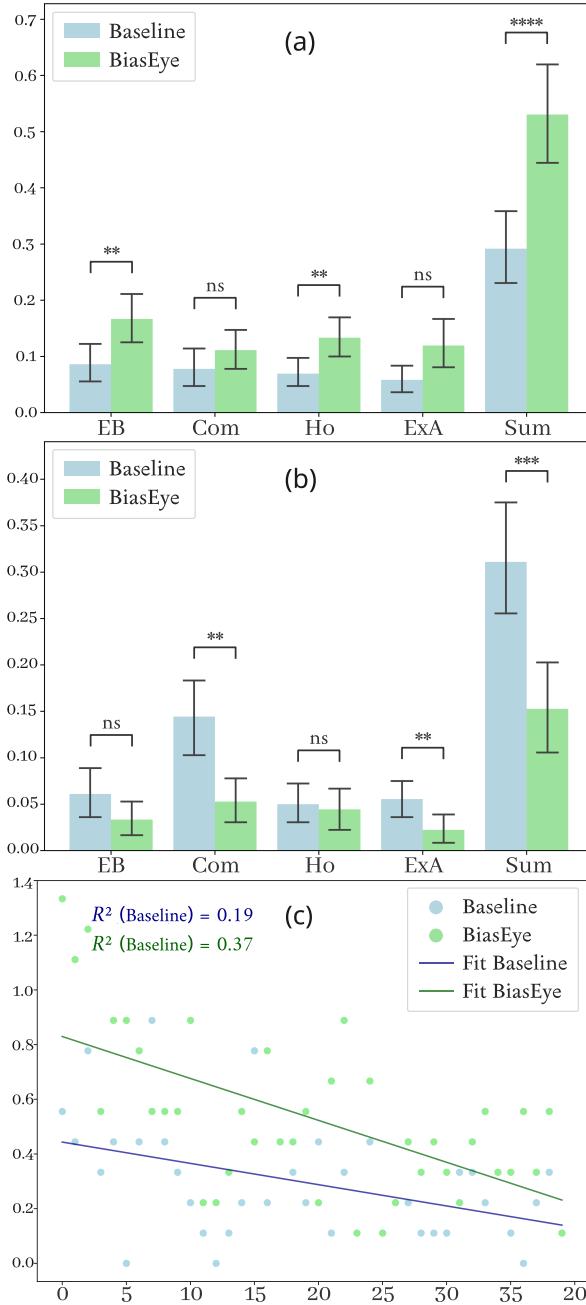


Figure 9: Differences in Revision Behavior Among Participants in Different Groups. (a) Differences in revision behavior during Phase I. (b) Differences in revision behavior during Phase II. (c) The average number of score modifications varies throughout the screening process in Phase I, with the horizontal axis representing application ID. The error bars indicate standard errors. (ns: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$; ****: $p < .0001$)

The results from Phase II further substantiate the notion that the *BiasEye* system contributes to the decrease of inconsistent decisions. As illustrated in Figure 9 (b), participants utilizing the *BiasEye* system exhibited significantly lower frequencies of revisions in the Com ($U = 1,086, p < 0.01$), ExA ($U = 1,028, p < 0.01$), and Sum ($U = 1,181, p < 0.001$) categories. Despite being exposed to more comprehensive global information in Phase II, participants employing the *BiasEye* system had already mitigate most of inconsistent decision outcomes during Phase I, decrease the requirement for additional score revisions. P10 explicitly stated, “without those charts on the left (Statistical View), it can be kind of hard for me to tell the difference between the different application levels because the scores start to blur together. Having those charts really makes a difference for me.”

Finding 8: Participants utilizing the *BiasEye* system exhibit more concentrated scoring for high-quality applicants, resulting in fewer instances of inconsistent outcomes. While cognitive bias can play a role, it’s important to recognize that different reviewers may hold varying opinions about an application. Existing literature, as mentioned in Coleman et al. [14], emphasizes the use of “interrater reliability” to ensure the effectiveness and consistency of screening decisions. One way to assess this is through “composite reliability”, as outlined by Coleman and colleagues [14], where a group of reviewers score within an acceptable range.

To evaluate whether the *Statistical* view in the *BiasEye* system helps mitigate screening inconsistencies, we compared the screening outcomes for high-quality applications in Phase I at different score levels (assuming equal section weights) in both Group A and Group B. We took the intersection of the results from both groups to ensure consistency. The outcomes are presented in Figure 10, where each bar represents the number of applications receiving a specific score. Here, we denote the number of compared applications as N and measure the kurtosis of the histograms as K .

Our observations indicate that Group B, using the *BiasEye* system with the *Statistical* view, exhibits more centralized outcomes, reflected in the higher kurtosis value (K). A similar trend is observed when comparing Phase I to Phase II, regardless of whether the Baseline or *BiasEye* system was used. This evaluation underscores the effectiveness of the *Summary* page in mitigating inconsistencies, as seen in Figure 11. It’s worth noting that the phenomenon in Figure 11 is less pronounced due to the comparison being based on an intersection, which excludes a significant portion of adjusted applications in Phase II.

6.2 RQ2. How will participants interact with and be affected by the bias-aware system in material screening?

Building upon the earlier-discussed analysis methods outlined in section 6, we initially explore the ways in which participants will engage with our bias-aware designs to fine-tune their decision outcomes. Subsequently, we proceed to unveil the discoveries regarding how these designs impact participants’ cognitive workload and their self-evaluation of the decisions made.

6.2.1 Usage pattern. This section presents the observations regarding how participants utilize our system design in a systematic

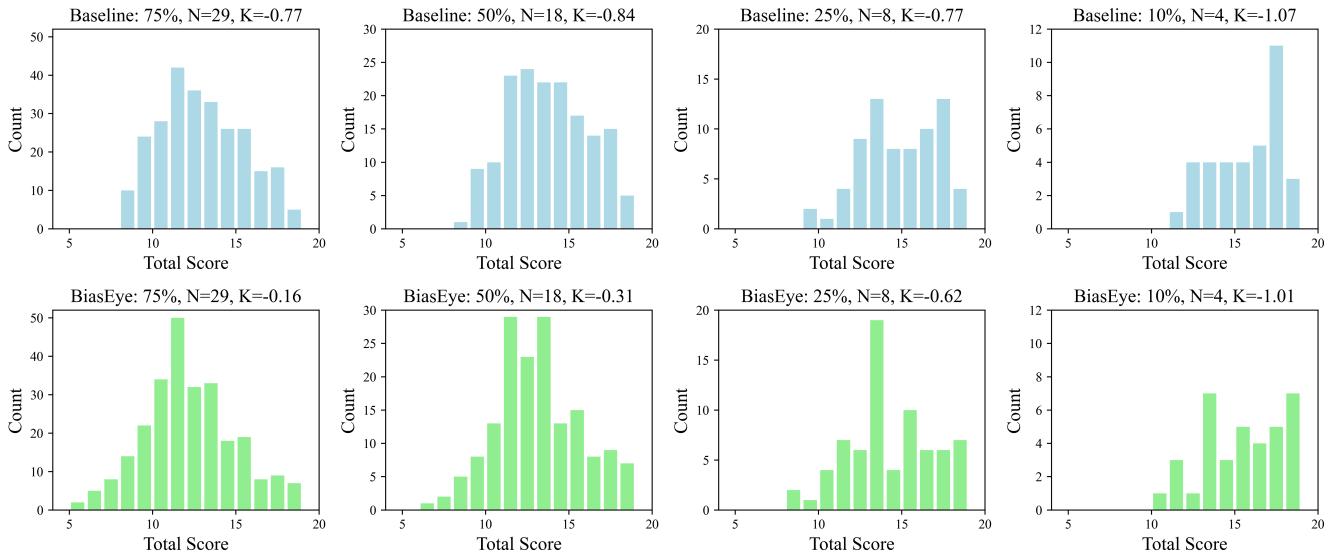


Figure 10: Differences in the score distribution between Baseline and BiasEye systems in Phase I.

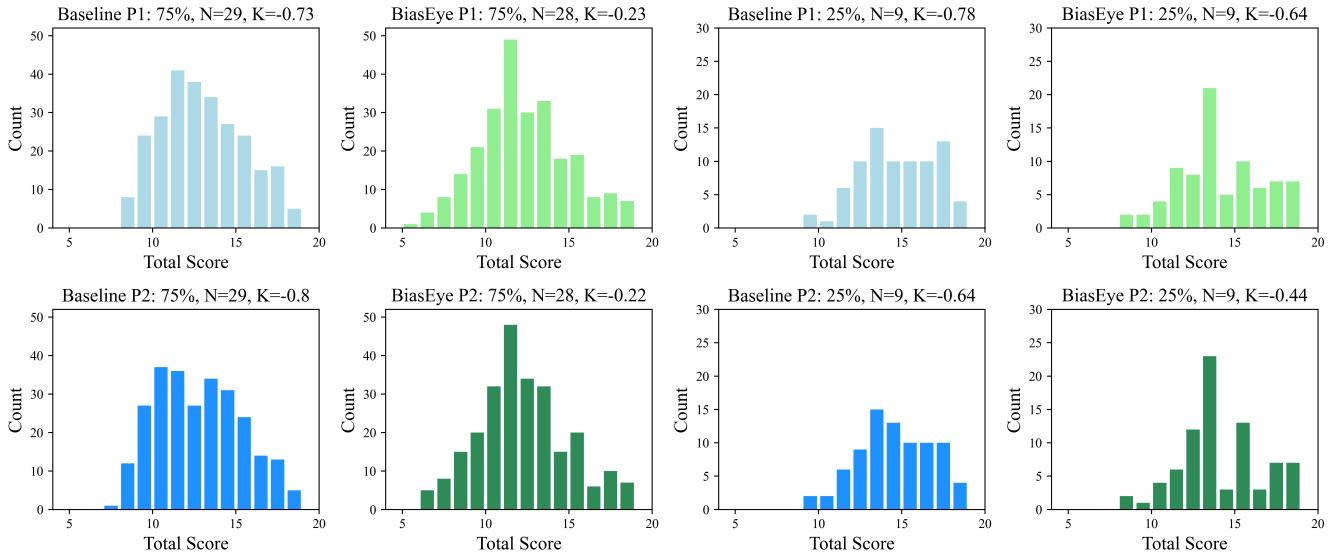


Figure 11: Differences in the score distribution between Phase I (P1) and Phase II (P2).

four-step process aimed at preventing, discovering, locating, and mitigating inconsistent decision outcomes.

Step 1. Preventing. Video transcription shows all Group B participants employed the *Statistical View* for insights into application materials and positioning applicants, and considered publication level as a screening criterion. Conversely, most Group A participants (7 out of 10) tried to check the given reference, but nearly all (6 out of 7) stopped after about 20 applications. Interestingly, three participants ignored this feature. These observations support our design motivation to enhance information transparency and accessibility.

Step 2. Discovering. Participants predominantly employ two categories of methods to identify inconsistencies in their decision outcomes on the *Summary* page. First, a minority of participants (3

out of 19) inspected exceptions to the time allocation in the *Ex-situ Table*. Second, the majority of participants (17 out of 19) used the back-end model to aid them in discovering potential anomalies through prediction scores. They selected trusted samples for back-end training through three distinct approaches:

- Most participants (13 out of 19, including 7 from Group A) directly chose applicants falling within a specific range based on the screening order. This range deliberately excluded the initial 5-10 applicants, as participants perceived their screening criteria to be either more lenient or stricter for this subset. This observation suggests that cognitive bias cannot be entirely eliminated, even with the aid of statistical information and supplementary.

- A subset of participants (5 out of 19, including 2 from Group A) manually selected representative applicants from each score category (1-5) using checkboxes.
- Participant P9 employed an unconventional approach that exceeded our expectations in sample selection. Initially, P9 included all applicants in the first round of training and then chose applicants exhibiting consistency between the model's predicted scores and human scores as the final samples for the second round of training. In the video, P9 mentioned being unfamiliar with machine learning but believed this approach could help identify samples that could serve as representatives of his scoring criteria. We observed an increase in the number of consistent outcomes after the second round of training, although this may have occurred by chance. Exploring whether repeating such operations could lead to convergence and automate the process is an intriguing topic for future research.

Step 3. Locating. Building upon the methods outlined in Step 2, participants employed specific strategies to identify anomalies. This process can be categorized into two distinct approaches. First, a minority of participants (3 out of 19) directly scrutinized applicants who received either inadequate or excessive time allocations, classifying them as cases of oversight or difficulty in decision-making, respectively. Second, participants who utilized the back-end model (comprising 17 out of 19) employed two primary methods to pinpoint applicants with potential inconsistencies:

- Five participants harnessed the sorting function within the 'Ex-situ Table'. They initially sorted the table based on the columns labeled 'EB/Com/Ho/ExA' or 'Mitigate'. Their focus was directed towards applicants where the order of predictions/human scores contradicted the ascending or descending order of human scores/predictions.
- Twelve participants identified potential inconsistencies by observing the ID color and assessing the variance between the two rings of a glyph. When confronted with multiple anomalies marked with blue or red colors, participants developed distinct patterns of focus: i) A majority (7 out of 12) concentrated on applicants displaying a high discrepancy between the two rings, a preference influenced by their personal perception. ii) Two participants focused on identifying the lower/higher scores within an overall trend of higher/lower scores. iii) Three participants searched for inconsistencies within the pool of applicants who had received high human scores.

These patterns of focus shed light on participants' expectations of generating rational screening outcomes.

Step 4. Mitigating. Participants accessed the *Screening Sheet* of the corresponding student on the *Summary* page by clicking on rows within the *Ex-situ Table*. They employed various strategies to adjust the assigned scores, which included: (1) Comparing the applicant's score with those who received the same score; (2) Comparing the applicant's score with individuals who had similar predicted scores from the model; (3) Comparing the applicant's score with students positioned closely in the *Comparison* view. (4) Relying entirely on, or taking into consideration, the model's recommendations; (5) Referring to the keywords listed in the notification card

to understand the model's rationale and checking if any relevant features were overlooked during Phase I; (6) Assessing the model's performance based on keywords and the distribution of ID colors in the *Comparison View* to determine whether further examination of potentially inconsistent applications was necessary. The sixth strategy is particularly relevant to the issue of trust in the model, and our findings related to this are presented in subsection 6.3. These strategies underscore the adaptability of our system design, accommodating the diverse usage habits and preferences of individual users while achieving the goal of mitigating inconsistent decision outcomes.

6.2.2 Effects on participants' cognitive workload. In this section, we employ questionnaire data to assess the variations in workload among participants when comparing the Baseline and *BiasEye* systems. The results are visually presented in Figure 12 (a). During Phase 1, *BiasEye* significantly reduced psychological ($U = 70.0, p < 0.01$) and time workload ($U = 72.0, p < 0.01$) compared to the Baseline. Transitioning from Phase 1 to Phase 2, the introduction of *Summary* page resulted in significant reductions in both psychological ($T = 0.0, p < 0.05$ in Baseline, $T = 0.0, p < 0.05$ in *BiasEye*) and physical workloads ($T = 0.0, p < 0.05$ in Baseline, $T = 5.5, p < 0.05$ in *BiasEye*) for both systems. Participants did not report significant changes about time workload and feeling of frustration.

6.2.3 Effects on participants' self-evaluation. Figure 12(b), we present the differences in self-evaluation between the Baseline and the *BiasEye* system. During Phase 1 of the experiment, participants using *BiasEye* reported experiencing more consistent criteria ($U = 8.0, p < 0.01$) and better distinction among applications ($U = 11.5, p < 0.01$) compared to the Baseline group. Additionally, *BiasEye* significantly improved the screening efficiency ($U = 6.0, p < 0.01$). Moving on to Phase 2, the introduction of the *Summary* page had a notable impact on both groups. It enhanced the criteria consistency ($T = 0.0, p < 0.05$ in the Baseline group and $T = 3.0, p < 0.05$ in the *BiasEye* group) and improved the distinction among applications ($T = 0.0, p < 0.05$ in the Baseline group and $T = 0.0, p < 0.05$ in the *BiasEye* group). However, it's important to note that only the Baseline group reported a significant increase in efficiency.

6.3 RQ3. How will participants trust and collaborate with the ML method?

Through a qualitative analysis of video transcripts, we identified varying levels of trust among participants in the suggestions provided by the model. This trust, in turn, influenced their collaborative interactions with the machine learning-supported assistant system. Among the 19 participants in our study, only two opted not to utilize the model. The remaining 17 participants all made revisions based on the model's recommendations. It's important to note that participants retained ultimate decision-making authority when it came to screening results. They determined whether to accept, refer to, or question the prediction scores of an application, integrating their own understanding of the application materials. The machine learning method served as a supplementary tool, offering a clear and expedited path to identify inconsistencies within specific applications. More specifically, our study revealed the following findings into participants' trust in and collaboration with the ML method.

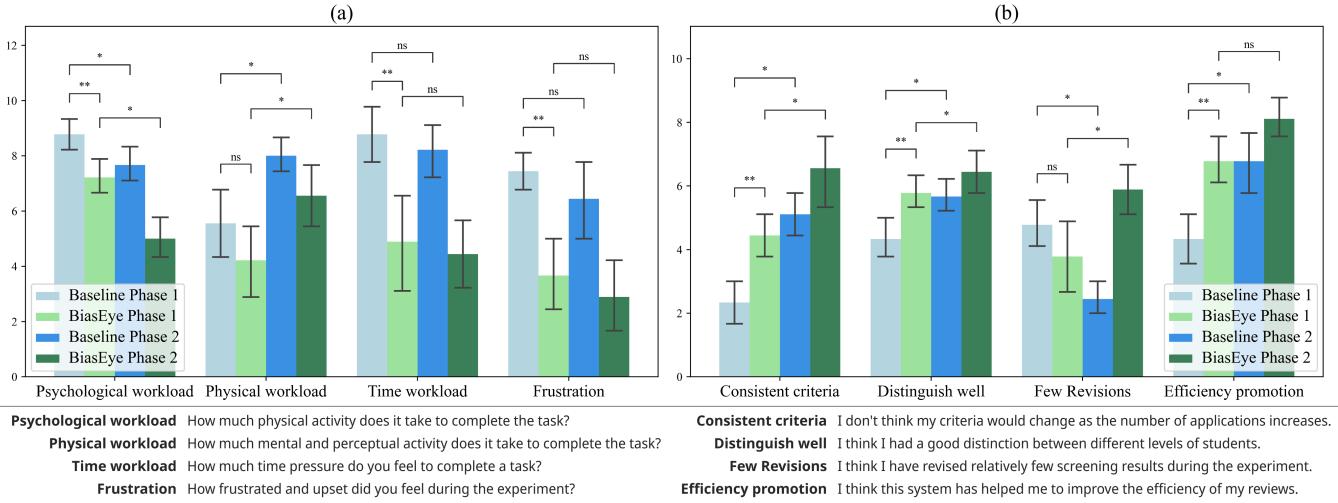


Figure 12: Results of the (a) workload assessment and (b) self-evaluation in the questionnaire. Error bars indicate standard errors. (ns: p < .1; *: p < .05; **: p < .01).

Finding 9: Participants' trust in the model's performance is screening section-independent. Participants' lack of trust in the model's performance in one section did not influence their trust in other sections. For instance, P6 remarked, “*The model's predictions in the Ho section are not accurate, but it does help me identify many incorrect scores in the EB section.*” Similarly, P11 expressed, “*I'm quite confident in how the model handles quantitative data, but the content in the ExA section encompasses various elements, and I doubt the model could comprehend my criteria.*”

Finding 10: Participants generally attempted to comprehend the rationale behind model predictions, but success was not guaranteed. Participants often inferred the reasons behind the model's predictions by examining various factors, including the prediction itself, attributes with significant weights displayed on the *Notification Card*, and raw information about each student. These inferences ranged from grasping the overall logical reasoning of the model to providing individual explanations for specific application predictions. For example, P3 remarked, “*Attributes on the notification involve scores of the English proficiency test (CET) and school ranking, but the model thinks I gave high scores for many applications... um, it is sensitive to CET scores, I care less unless the CET score is under 500.*” Conversely, according to P13, “*The model scores 2, but he comes from an experimental class at a university, with understandably low ranking that the model might have overlooked. I'm sticking to my opinion.*” As the system did not explicitly specify the concrete attributes contributing to each application's prediction, there were instances where participants found it challenging to make successful inferences, leading to comments such as, “*I can't understand,*” as noted by several participants.

Finding 11: Participants tended to question the rationality of their decisions when there was a significant disparity between the predictions and human scores/expectations. Participants' awareness of these differences stemmed from two primary sources. On one hand, it was influenced by the overall color trend of the ID text in the *Comparison view*, as noted by P16 who mentioned, “*There are many red colors, and I am overwhelmed.*” On the other hand, participants observed discrepancies between

the human scores they assigned and the predicted scores for each application. P8 remarked, “*The prediction is around 5 points, but why did I only give 1 point? Although I don't quite understand why it scores 5, I decide to increase the score a bit.*” Despite being informed at the beginning of the experiment that the model's predictions may be inaccurate, participants still exhibited a degree of blind belief and reliance on the model, particularly when they felt uncertain about an application. As P8 questioned, “*I am struggling with this score, or should I listen to the model?*” Nevertheless, it's worth noting that the proposed system has mitigated confirmation bias to some extent by encouraging participants to engage in a second round of deliberation.

Through an analysis of the video transcripts, we also identified various factors that influenced participants' trust.

Factor 1: The consistency between keywords and participants' perception of decision criteria. The attributes listed on the *Notification Card* served as the initial point for participants to grasp the model's functioning. When these attributes did not align with the participants' preconceived criteria, it led to doubts regarding the model's predictions. For instance, P11 exhibited skepticism towards the attributes in the *ExA* section. Upon identifying discrepancies and disagreeing with the predictions for three applications in the *Comparison view*, P11 promptly cross-referenced and verified the scores of multiple applications in the *Ex-situ Table* independently. We observed that inconsistent perceptions could also arise from differences in how attributes were categorized and participants' mental frameworks. For instance, competitions were initially categorized into different subjects within the *Com* section. However, participants were often unaware of the distinctions between different subject areas within competitions, particularly for competitions that were rarely mentioned in the materials and thus not well-remembered or paid attention to.

Factor 2: The significance of differences between predictions and human scores. Participants exhibited strong trust in prediction scores that closely matched or were consistent with human scores. None of the participants actively sought applications

with gray-colored IDs in the *Comparison* view or those where human scores and predictions were sorted in the same order in the *Ex-situ Table*. As P9 stated, “*Both the model and I agree with these scores, so there’s no issue at all.*” The greater the difference between the human score and prediction, the more likely it was for inconsistencies in applications to exceed the participants’ threshold and capture their attention. For instance, P6 commented, “*Differences less than one are not an issue. I’ll check the others that had larger score differences.*” Conversely, the overall color trend of IDs in the *Comparison* view also influenced participants’ trust in the model’s predictions. P12 mentioned, “*There are many gray ones, so I believe that the model has learned well.*” It’s essential to note that this factor does not contradict the phenomenon of self-doubt arising from higher/lower color trends mentioned in Finding 11.

Factor 3: The presence of sufficient evidence for confirmation and trust. Participants were more inclined to trust predictions when they discovered ample evidence to support them. This evidence could be gathered by verifying whether key information in an application had been overlooked or by making comparisons between multiple applications. For example, P4 admitted, “*It’s my fault. I didn’t pay attention to the Mathematical Contest In Modeling just now.*” In a similar vein, P11 commented, “*Compared to other applications that meet my expectations of four, this application is indeed slightly worse. I will follow the model and adjust it to three.*” Differences in how participants and the model interpreted the same information sometimes hindered their trust in the predictions. For example, P13 from Group A, which did not have access to the *Statistical* view, questioned, “*Why did the model give him a score of four when he’s from an average university and his GPA is not at the top level?*” Subsequently, the participant referred to supplementary materials and found that the university was ranked around the top 50, which is considered quite good. This incident highlights how human judgment can be influenced by personal experiences, potentially leading to biases, such as the availability heuristic [58] and confirmation bias, which makes individuals ignore objective truths. Notably, this issue was not observed among participants in Group B (*BiasEye*), as the *Statistical* view provided valuable evidence.

Factor 4: Participants’ intrinsic perceptions of machine learning. Participants’ intrinsic beliefs about machine learning significantly influenced their trust in the system. For instance, P7 expressed confidence, stating, “*The system is definitely more accurate than I am.*” Similarly, P15 held the view that, “*Machines don’t get tired; they have no blind spots in attention.*” Conversely, some participants like P4 were more skeptical, stating, “*I’ve learned about machine learning algorithms. If some attributes do not appear in the selected samples, it cannot learn them.*” P18 also struck a balance, noting, “*I believe that machine learning can assist me, but I’m aware it has limitations too. I won’t blindly follow it.*” These inherent perceptions of machine learning played a pivotal role in shaping participants’ trust.

7 DISCUSSION AND LIMITATION

In this section, we extract future design considerations DC1~4 (subsection 7.1) from our analysis results and questionnaire feedback. We also explore potential generalizations of our findings to other domains in subsection 7.2 and reflect on the limitations of our work in subsection 7.3.

7.1 Design Consideration

DC1: Improve the interactive capability of the system. Participants appreciated *BiasEye*’s interactive features in our study, such as real-time score box-plot updates, highlighting the current application in the *Statistical* view, and quick navigation between *Screening Sheets*, which alleviate their workload. A bias-aware intelligent interface for decision-making should seamlessly incorporate interactive functionality, enabling users to devote more cognitive resources to thoughtful judgment. This integration is essential for encouraging users to actively address biases in decision outcomes. Additionally, such systems should gather and present more contextual information to support well-informed decisions. Our study revealed that certain participants in group A, like P5 and P13, infrequently referred to supplementary materials and were influenced by personal experiences, leading to inconsistent screening results. To alleviate the impact of inadequate or incorrect memory and perception, a recommendation is to implement dynamic annotations within the interface. These annotations could include hyperlinks to pertinent information such as school, major, competition details, and data on past admitted students. If this information could be aggregated, the interface might visualize a comparison between individual and collective data. Consequently, instead of facing unfamiliar and ambiguous perceptions, users could swiftly grasp relevant information.

DC2: Simplify views and visual designs. The design of visualization and functionality should prioritize intuitiveness, avoiding the need for complex computer expertise and minimizing the learning curve. In our study, participants acknowledged the attractiveness of glyphs but found their placement lacked meaning, as highlighted by P6 and P11. The process of visualized dimensionality reduction added cognitive demands and had the potential to cause misunderstanding. Interestingly, the *Ex-situ Table* view was deemed more user-friendly than the *Comparison* view, leading participants to prefer a format combining glyphs with a table presentation. As a result, future interface designs could incorporate tables with multiple straightforward mini-charts, offering a more effective way for users to understand data without increasing cognitive load. Additionally, for complex decision tasks like material screening, it remains uncertain whether a multi-view visual analysis strategy is a more effective option.

DC3: Enhance machine learning with human guidance. Our observations unveiled that pre-specified model training attributes approximated only a limited subset of participants’ personal criteria. Despite some commonalities, each participant had unique focus areas. A universal model struggled to differentiate individual applications based on specific criteria and often misclassified similar applications due to attribute redundancy. While more intricate models and comprehensive attributes could align better with actual screening criteria, there exists a trade-off between a perfect fit and real-time response. AI methods may not be as proficient or accurate as domain experts in verifying applicants’ contributions and identifying potential exaggerations. Moreover, AI faces challenges in acquiring contextual knowledge, such as how personal experiences are influenced by socioeconomic and geographic disparities. Implicit discrimination may be hidden in the superficial quantification of applicants based on factors like SAT scores and

academic awards. To address the limitations of ML methods, future intelligent screening systems should adopt “human-in-the-loop” approaches. Specifically, the interface can allow model training for customized attributes, correction of deviant model, special marking and score lock of outliers (e.g., students at risk of fraud or those considered deserving of preferential treatment).

DC4: Acknowledge the constraints of AI assistance techniques. The majority of participants (11 out of 18, with 5 not providing a response) acknowledged that automated information extraction improved retrieval efficiency and reduced their workload. Additionally, they found that the ML method assisted in addressing inconsistencies in screening decisions. However, it was also observed that participants tended to heavily rely on AI support methods, particularly the ML predictions. Consider the limitations of ML methods mentioned in DC3, human-machine collaboration strategies should be devised to promote AI in complement with human decision-making, rather than allowing unchecked dependence on algorithms. In this context, future intelligent interfaces should discourage the outright use of AI in initial decision-making, instead supervising users to adopt recommendations with adequately understanding. For example, system can pop up temporary windows to declare the limitation of the AI method, inquire about users’ confidence in their personal judgment versus AI prediction, and encourage users to assess the consistency of their judgment with AI recommendations.

7.2 Generalizability

Tasks such as corporate hiring, fund applications, and scholarship selections often require the evaluation of numerous multi-dimensional and multi-modal materials. These tasks commonly face different cognitive biases, resulting in inconsistent outcomes and affecting individual fairness. *BiasEye* is flexible and can be tailored by modifying the necessary attributes and algorithms to meet the specific requirements of a task. In our user studies, the simple *Ranking SVM* demonstrated encouraging results in assisting with bias mitigation. We are also interested in exploring more advanced approaches, such as neural networks, capable of capturing complex reasoning processes to further improve the effectiveness of bias mitigation.

7.3 Limitation

This study primarily evaluates our bias-aware screening system design, excluding information extraction as an experimental condition. However, it’s important to note that data extraction and classification models can introduce errors, highlighting the need for better document organization in application submissions. We recommend institutions implement formal systems for collecting structured personal information alongside documents, which can improve screening system design and functionality. Additionally, *BiasEye* relied on quantitative attributes for prediction, potentially missing nuanced human screening preferences, especially for indicators like project content and quality. To address this, exploring specialized language models or textual information extraction features may enhance learning and prediction, particularly in detecting biases in PSs and LoRs. Future systems could also simplify screening through content analysis for categorical comparisons of applicants. Lastly, due to constraints, we conducted a controlled in-lab study

with senior students, not directly comparable to expert admissions reviewers. We plan to pursue a field study after further system optimizations.

8 CONCLUSION AND FUTURE WORK

This study introduces *BiasEye*, a specialized interactive system designed to address, detect, and mitigate potential biases in real-time screening processes. *BiasEye* provides users with clear global views of information, aiding in fair screening criteria formulation. It also helps identify biases by comparing actual rankings with model-predicted ones, offering immediate means for adjustment. Results from a user study show that *BiasEye* significantly improves reviewers’ decision-making by visualizing potential biases, suggesting its value across screening tasks. Future improvements may involve advanced machine learning algorithms and broader domain applications, including enterprise and government contexts. *BiasEye* development could inspire more tools for impartial decision-making and bias reduction.

ACKNOWLEDGMENTS

We would like to express our gratitude to our domain experts and the anonymous reviewers for their insightful comments. This work is funded by grants from the National Natural Science Foundation of China (No. 62372298), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), and the Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Ministry of Education.

REFERENCES

- [1] Naeem Akl and Ahmed Tewfik. 2016. Designing interventions to mitigate cognitive biases in human decisions. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, Vietri sul Mare, Italy, 1–6. <https://doi.org/10.1109/MLSP.2016.7738838>
- [2] Naeem Akl and Ahmed Tewfik. 2016. Optimal information ordering for sequential detection with cognitive biases. In *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, Budapest, Hungary, 413–417. <https://doi.org/10.1109/EUSIPCO.2016.7760281>
- [3] M. Alamelu, D.Sathish Kumar, R. Sanjana, J.Subha Sree, A.Sangeerani Devi, and D. Kavitha. 2021. Resume Validation and Filtration using Natural Language Processing. In *2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IMECON)*. IEEE, Jaipur, India, 1–5. <https://doi.org/10.1109/IMECON53809.2021.9689075>
- [4] Association of American Medical Colleges. 2021. Holistic Review. <https://www.aamc.org/services/member-capacity-building/holistic-review>.
- [5] Jiang Bian, Jason Greenberg, Jizhen Li, and Yanbo Wang. 2022. Good to Go First? Position Effects in Expert Evaluation of Early-Stage Ventures. *Manage. Sci.* 68, 1 (jan 2022), 300–315. <https://doi.org/10.1287/mnsc.2021.4132>
- [6] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D³ Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309. <https://doi.org/10.1109/TVCG.2011.185>
- [7] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.
- [8] Alafair S Burke. 2005. Improving prosecutorial decision making: Some lessons of cognitive science. *Wm. & Mary L. Rev.* 47 (2005), 1587.
- [9] Kathleen Cachel, Elke Rundensteiner, and Lane Harrison. 2022. MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, Kuala Lumpur, Malaysia, 1124–1137. <https://doi.org/10.1109/ICDE53745.2022.00089>
- [10] 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>
- [11] Alexander Chernev, Ulf Böckenholdt, and Joseph Goodman. 2015. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology* 25, 2 (2015), 333–358. <https://doi.org/10.1016/j.jcps.2014.08.002>
- [12] 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, Phoenix, AZ, USA, 116–126. <https://doi.org/10.1109/VAST.2017.8585665>
- [13] Tee Chuanromane and Ronald Metoyer. 2022. A Crowdsourced Study of Visual Strategies for Mitigating Confirmation Bias. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Roma, Italy, 1–6. <https://doi.org/10.1109/VL/HCC53370.2022.9833151>
- [14] Arthur L Coleman and Jamie Lewis Keith. 2018. Understanding holistic review in higher education admissions. *New York: College Board* (2018).
- [15] Pat Croskerry, Geeta Singhal, and Silvia Mamede. 2013. Cognitive debiasing 2: impediments to and strategies for change. *BMJ Quality & Safety* 22, Suppl 2 (2013), ii65–ii72. <https://doi.org/10.1136/bmjqqs-2012-001713>
- [16] E. Deros and A. M. Ryan. 2018. By any other name: discrimination in resume screening. (2018). <https://doi.org/10.1093/oxfordhb/9780199764921.013.017>
- [17] Eva Deros and Ann Marie Ryan. 2019. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal* 29, 2 (2019), 113–130. <https://doi.org/10.1111/1748-8583.12217>
- [18] Ketki V. Deshpande, Shimei Pan, and James R. Fouls. 2020. Mitigating Demographic Bias in AI-Based Resume Filtering (*UMAP '20 Adjunct*). Association for Computing Machinery, New York, NY, USA, 268–275. <https://doi.org/10.1145/3386392.3399569>
- [19] Evanthis Dimara, Gilles Bailly, Anastasia Bezerianos, and Steven Franconeri. 2019. Mitigating the Attraction Effect with Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 850–860. <https://doi.org/10.1109/TVCG.2018.2865233>
- [20] Evanthis Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2020. A Task-Based Taxonomy of Cognitive Biases for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 26, 2 (2020), 1413–1432. <https://doi.org/10.1109/TVCG.2018.2872577>
- [21] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-computer interaction*. Pearson Education. 2579–2605 pages. https://doi.org/10.1007/978-0-387-39940-9_192
- [22] Jessica Marin Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 161, 9 pages. <https://doi.org/10.1145/3491102.3517443>
- [23] Baruch Fischhoff. 1975. The Silly Certainty of Hindsight. *Psychology today* 8, 11 (1975), 70.
- [24] Joseph P Forgas and Simon M Laham. 2016. Halo effects. In *Cognitive illusions*. Psychology Press, 286–300.
- [25] Bodhvi Gaur, Gurpreet Singh Saluja, Hamsa Bharathi Sivakumar, and Sanjay Singh. 2021. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. *Neural Computing and Applications* 33, 11 (2021), 5705–5718. <https://doi.org/10.1007/s00521-020-05351-2>
- [26] Juan E. Gilbert. 2021. Equitable AI: Using AI to Achieve Diversity in Admissions. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) (IUI '21). Association for Computing Machinery, New York, NY, USA, 1. <https://doi.org/10.1145/3397481.3457410>
- [27] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. sage publications.
- [28] Tumula Mani Harsha, Gangaraju Sai Moukthika, Dudipalli Siva Sai, Mannuru Naga Rajeswari Pravallika, Satish Anamalamudi, and MuraliKrishna Enduri. 2022. Automated Resume Screener using Natural Language Processing(NLP). In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, Tirunelveli, India, 1772–1777. <https://doi.org/10.1109/ICOEI53556.2022.9777194>
- [29] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*. Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [30] Yaozong Jia and Xiaobin Xu. 2018. Chinese Named Entity Recognition Based on CNN-BiLSTM-CRF. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, Beijing, China, 1–4. <https://doi.org/10.1109/ICSESS.2018.8663820>
- [31] Christel Joubert, Charlene Downing, and Irene J. Kearns. 2022. Selection process for admission to an academic nursing programme – A meta-synthesis. *Nurse Education Today* 116 (2022), 105475. <https://doi.org/10.1016/j.nedt.2022.105475>
- [32] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
- [33] R. Klinkenberg and T. Joachims. 2000. Detecting Concept Drift with Support Vector Machines. In *International Conference on Machine Learning (ICML)*. Morgan Kaufmann, San Francisco, 487–494.
- [34] Vivian Lai, Kyong Jin Shim, Richard J. Oentaryo, Philips K. Prasetyo, Casey Vu, Ee-Peng Lim, and David Lo. 2016. CareerMapper: An automated resume evaluation tool. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington, DC, USA, 4005–4007. <https://doi.org/10.1109/BigData.2016.7841091>
- [35] XiaoWei Li, Hui Shu, Yi Zhai, and ZhiQiang Lin. 2021. A Method for Resume Information Extraction Using BERT-BiLSTM-CRF. In *2021 IEEE 21st International Conference on Communication Technology (ICCT)*. IEEE, Tianjin, China, 1437–1442. <https://doi.org/10.1109/ICCT52962.2021.9657937>
- [36] Jerome A Lucido. 2014. How Admission Decisions Get Made. *Handbook of Strategic Enrollment Management* 1 (2014), 158–173.
- [37] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
- [38] Liz Melton and Grant Riewe. 2022. Using AI to minimise bias in an employee performance review. *Journal of AI, Robotics & Workplace Automation* 2, 1 (2022), 17–23.
- [39] Ronald A. Metoyer, Tee Chuanromane, Gina M. Girgis, Qiyu Zhi, and Eleanor C. Kinyon. 2020. Supporting Storytelling With Evidence in Holistic Review Processes: A Participatory Design Approach. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 61 (may 2020), 24 pages. <https://doi.org/10.1145/3392870>
- [40] 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>
- [41] Sara Nasr and Oleg Vitoldovitz German. 2019. Assessment of Graduate Students' Resumes Using Short Text Searching Method. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, Sardinia, Italy, 306–308. <https://doi.org/10.1109/AIKE.2019.00061>
- [42] S. M. Noble, L. L. Foster, and S. B. Craig. 2021. The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment* (2021). <https://doi.org/10.1111/ijsa.12320>
- [43] Alexander Nussbaumer, Katrien Verbert, Eva-Catherine Hillemann, Michael A. Bedek, and Dietrich Albert. 2016. A Framework for Cognitive Bias Detection and Feedback in a Visual Analytics Environment. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, Uppsala, Sweden, 148–151. <https://doi.org/10.1109/EISIC.2016.038>
- [44] Alejandro Peña, Ignacio Serna, Aynthami Morales, and Julian Fierrez. 2020. Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 129–137. <https://doi.org/10.1109/CVPRW50498.2020.00022>
- [45] Grant A Pignattiello, Richard J Martin, and Ronald L Hickman Jr. 2020. Decision fatigue: A conceptual analysis. *Journal of health psychology* 25, 1 (2020), 123–135. <https://doi.org/10.1177/1359105318763510>
- [46] Margit Pohl, Lisa-Christina Winter, Chris Pallaris, Simon Attfield, and B. L. William Wong. 2014. Sensemaking and Cognitive Bias Mitigation in Visual Analytics. In *2014 IEEE Joint Intelligence and Security Informatics Conference*. IEEE, The Hague, Netherlands, 323–323. <https://doi.org/10.1109/JISIC.2014.68>
- [47] Rüdiger Pohl. 2016. *Cognitive illusions: Intriguing phenomena in thinking, Judgment and Memory*.
- [48] Rasika Ransing, Akshaya Mohan, Nikita Bhrugumaharshi Emberi, and Kailas Mahavarkar. 2021. Screening and Ranking Resumes using Stacked Model. In *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECOT)*. IEEE, Mysuru, India, 643–648. <https://doi.org/10.1109/ICEECOT52851.2021.9707977>
- [49] Mohammad Selim. 2021. Anchoring bias in Corporate decision making and its effects on net income. In *2021 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, Sakheer, Bahrain, 424–429. <https://doi.org/10.1109/DASA53625.2021.9682369>
- [50] Uri Simonsohn and Francesca Gino. 2013. Daily horizons: Evidence of narrow bracketing in judgment from 10 years of MBA admissions interviews. *Psychological science* 24, 2 (2013), 219–224. <https://doi.org/10.1177/0956797612459762>
- [51] Atanu R Sinha, Navita Goyal, Sunny Dhammani, Tanay Asija, Raja K Dubey, MV Raja, and Georgios Theocaras. 2022. Personalized Detection of Cognitive Biases in Actions of Users from Their Logs: Anchoring and Recency Biases. (2022). <https://doi.org/10.48550/arXiv.2206.15129>
- [52] Submittable. 2021. Holistic Review: What It Is and How to Use It. <https:////blog.submittable.com/holistic-review/>.
- [53] Poorna Talkad Sukumar and Ronald Metoyer. 2018. A visualization approach to addressing reviewer bias in holistic college admissions. *Cognitive biases in visualizations* (2018), 161–175.
- [54] Poorna Talkad Sukumar, Ronald Metoyer, and Shuai He. 2017. Holistic Reviews in Admissions: Reviewer Biases and Visualization Strategies to Mitigate Them. In *IEEE VIS 2017*. IEEE, 5 pages.
- [55] Pratibha Swami and Vibha Pratap. 2022. Resume Classifier and Summarizer. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Vol. 1. IEEE, Faridabad, India, 220–224. <https://doi.org/10.1109/COM-IT-CON54601.2022.9850527>
- [56] Poorna Talkad Sukumar, Ronald Metoyer, and Shuai He. 2018. Making a Pecan Pie: Understanding and Supporting The Holistic Review Process in Admissions. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 169 (nov 2018), 22 pages. <https://doi.org/10.1145/3274438>
- [57] B Alden Thresher. 1966. College Admissions and the Public Interest. (1966), 132 pages.

- [58] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- [59] André Calero Valdez, Martina Ziefle, and Michael Sedlmaier. 2018. Priming and Anchoring Effects in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 584–594. <https://doi.org/10.1109/TVCG.2017.2744138>
- [60] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008)
- [61] Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert. 2019. A Formative Study of Interactive Bias Metrics in Visual Analytics Using Anchoring Bias. In *Human-Computer Interaction – INTERACT 2019 (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 555–575. https://doi.org/10.1007/978-3-030-29384-0_34
- [62] 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, Phoenix, AZ, USA, 104–115. <https://doi.org/10.1109/VAST.2017.8585669>
- [63] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T. Brown, and Alex Endert. 2018. Podium: Ranking Data Using Mixed-Initiative Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 288–297. <https://doi.org/10.1109/TVCG.2017.2745078>
- [64] Emily Wall, Arpit Narrechania, Adam Coscia, Jamal Paden, and Alex Endert. 2022. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 966–975. <https://doi.org/10.1109/TVCG.2021.3114862>
- [65] 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, Vancouver, BC, Canada, 111–115. <https://doi.org/10.1109/VISUAL.2019.8933611>
- [66] Frank Wilcoxon, SK Katti, Roberta A Wilcox, et al. 1970. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected tables in mathematical statistics* 1 (1970), 171–259.
- [67] Qian Zhu, Leo Yu-Ho Lo, Meng Xia, Zixin Chen, and Xiaojuan Ma. 2022. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31. <https://doi.org/10.1145/3555597>

Appendix

A BASELINE DESIGN

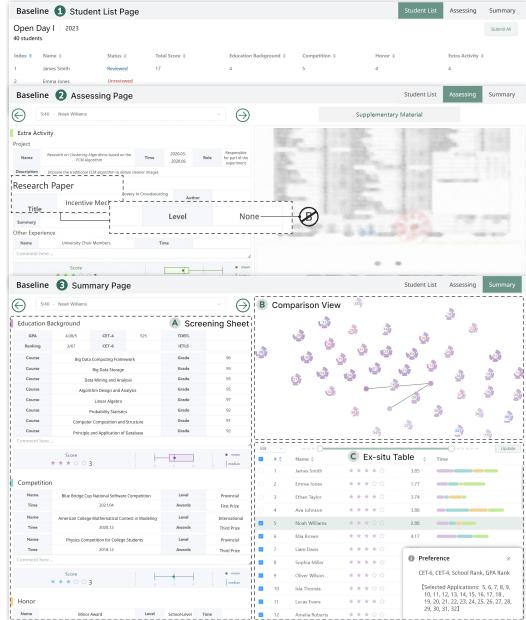


Figure 13: The baseline system used in the experiment. In Phase I, participants are limited to using only (1) the *Student List* page and (2) the *Assessing* page. In Phase II, participants can also utilize (3) the *Summary* page.

B USAGE SCENARIO

Let's consider Professor Alex, tasked with reviewing a large number of university admissions applications each May, faced the challenge of juggling this with his teaching and research duties. He found that meticulously reading through a large volume of application information consumed a significant amount of his energy. To cope with this, he resorted to manually extracting essential information from the PDF files, enabling him to make comparisons between applications before and after this data retrieval step, ultimately aiding in his decision-making process. While these heuristics proved effective in most cases, they did carry a risk of errors, especially when Alex began to experience fatigue. However, considering the significance of college admissions as a crucial societal matter, Alex dedicated a substantial amount of time to meticulously review his screening results before submission. He did this out of concern that he might overlook outstanding applicants in the process. This particular aspect of the procedure heavily relied on Alex's subjective judgment.

This year, Alex utilized *BiasEye* for the application screening process. He began by reviewing the *Statistical* view to conduct an initial evaluation of the applicants before commencing the screening. As the process continued, he increasingly relied on this view to recall and grasp quantitative information, such as school rankings. *BiasEye* conveniently extracted information according to sections in advance, effectively conserving Alex's energy. After evaluating approximately forty to fifty applications, Alex proceeded to the *Summary* page to examine his screening outcomes. Initially, he noticed some conspicuous irregularities in time allocation and identified an incorrect score in the *Extra Activity* section of one application. He realized that he had misread some vital information during his hurried decision-making process. Subsequently, Alex adopted a systematic approach to reviewing each section. He began by selecting the education section from the drop-down menu. Recognizing that he might have been too cautious initially due to a lack of confidence and that his decision-making quality had decreased towards the end of the process due to fatigue, he adjusted the slider to focus on the middle subset of applications that aligned with his screening preferences. He carefully inspected the contents on the *Notification Card* and found that the attributes aligned with his expectations. Alex noted that the *Comparison* view effectively highlighted discrepancies between the model and human 'decisions', making it easier to identify applications with lower or higher scores based on the color of the IDs. However, he felt that he should prioritize applications with a significant disparity between the inner and outer arcs. Additionally, by examining the center dots, he realized that he had only assigned ratings ranging from 2 to 5. While reviewing application #11, Alex observed that the score was lower than the prediction. In the *Comparison* view, he discovered that this application closely resembled #14, indicating a degree of similarity. Alex conducted a detailed comparison between these two applications and decided to follow the model's suggestion by modifying the score to 3. Employing a similar approach, he made adjustments to more scores. In the end, Alex's confidence in his screening results grew, and he was pleased to have mitigated the inconsistencies in his screening decisions.