Vɪs4GRAD: A Visual Analytic Approach for Promoting Process Awareness in Graduate Admissions

ANONYMOUS AUTHOR(S)

Graduate admissions is a complex decision making process where cognitive and implicit biases may impact the way reviewers individually and collectively make decisions. Education-based methods such as training courses have been applied to address such biases; however, since these biases are unconscious, informing reviewers about the existence of biases has limited impact during the decision making process. In this paper, we introduce a visualization system, Vɪs4GRAD, that promotes reviewers' self-reflection and scrutiny to ensure fair and consistent review processes. The visualization interface logs reviewers' interactions in order to provide a granular analysis of review behaviors across attributes such as race and gender of applicants which can relate to potentially biased processes and inform review procedures in subsequent cycles. We present results of a case study where the system was used for Ph.D. admissions in the Computer Science department at a private university. The results suggest that Vɪs4GRAD has potential to iteratively transform graduate review processes to ensure adherence to fair procedural goals e.g., to program goals for diversity, equity, and inclusion.

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

Additional Key Words and Phrases: process awareness, graduate admissions, decision making

ACM Reference Format:

Anonymous Author(s). 2022. Vɪs4GRAD: A Visual Analytic Approach for Promoting Process Awareness in Graduate Admissions. In . ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Graduate admissions involves complex decision making processes, often characterized by individual reviewers reviewing applications and rating them based on factors such as the applicant's academic performance, non-academic accomplishments, and personal qualities. Committee members' individual evaluations are collated and discussed to inform final admissions decisions. Due to the complexity and subjectivity of the decision-making process, **unconscious biases** might impact the way reviewers make admissions decisions. This work focuses on such unconscious biases including cognitive biases – systematic errors resulting from the use of simplifying heuristics when making judgments and decisions[20, 50] and implicit biases such as gender and racial bias that are ingrained as a result of a person's cultural beliefs and past experiences[22, 23]. Such biases are, by nature, difficult to perceive during the decision making process, and it is hard for reviewers to determine if a decision is biased without being aware of the factors that drive the decision making process. Our goal in this work is therefore to investigate (1) how we can promote real-time reflective review processes and (2) to what extent heightened awareness leads to reviewers' adjustment of associated behaviors and decisions.

Reviewer biases can lead to decisions that are both harmful to students and program missions. For example, one laboratory experiment showed that science faculty exhibit a bias against female students such that male candidates

Manuscript submitted to ACM

are rated as significantly more competent and hireable than (otherwise credentially identical) female candidates[38]. Similarly, racial discrimination has been studied in the labor market. Experimental results show significant discrimination against African American names in job applications: equivalent resumes with White names receive 50% more callbacks for interviews than resumes with African American names[6]. Inspired by prior work on mitigating biases in visual data analysis by visualizing users' interaction history with data to increase awareness of potential biases[40, 58] and reflective design [44], we designed an interactive system, **Vis4GRAD**, to support process-aware graduate admissions. We use the term **process awareness** to mean knowledge of what and how decisions have been made during the decision making process, which we aim to achieve by encouraging real-time **self-reflection** on review process through visualizations of interaction history. We focus in particular on promoting reviewers' self-assessment of internal consistency in their process across applications stratified by attributes such as race and gender, as measured by time spent and application ratings.

Vis4GRAD supports a general framework for admissions review, i.e., it enables admissions committee members to evaluate applications holistically and rate them along relevant factors defined by the admissions committee, followed by support for collaborative decision making with collated individual evaluations. Importantly, the system captures reviewers' interactions to compute time spent on each *applicant* and time spent on each *component of the application* (i.e., personal statement, resume, letters of recommendation, etc). Vis4GRAD provides visualizations of these interactions according to applicant attributes such as race and gender to promote reviewers' self-reflection of potentially biased behavior. Ambiguous review behaviors and decisions make it infeasible to detect each particular type of bias that may take place during the admissions review process. Instead we focus on increasing reviewers' awareness of behavioral indicators that might result from underlying biases. These behaviors could include not spending enough time on a certain application, systematically spending more/less time on a certain group of applications across gender or racial lines, and/or inconsistency in ratings among similarly qualified applicants, etc. These visualizations are available for each individual reviewer to reflect on their own processes, as well as in an aggregated format for the whole group to analyze collective trends in the committee's review process and decisions.

Our primary contributions include (1) the system, Vis4GRAD, designed in collaboration with two graduate program admissions committee chairs to promote reviewers' reflection on their review processes and increase awareness of undesired review behaviors, and (2) results of a case study with the admissions committee in the Computer Science department at a private university that demonstrates how Vis4GRAD can facilitate process-aware decision making. These results importantly showcase the real-world potential and possible ethical implications of process-aware visualization.

2 RELATED WORK

Relevant to the design and implementation of Vis4GRAD and execution of our case study, we discuss how increasing process awareness during university admissions can empower reviewers to address potential **biases in admissions** (Section 2.1), informed by prior approaches to **analytic provenance** (Section 2.2), **collaborative visualization** (Section 2.3), and recent work on **biases in visualization** (Section 2.4).

2.1 Biases in Admissions

A number of specific biases have been identified as particularly concerning in the context of university admissions. Implicit biases[23] in particular, shaped by cultural and societal norms (e.g., racial bias and gender bias), can perpetuate harmful societal stereotypes. The implicit Association Test (IAT)[24] characterizes such biases by measuring the association that people hold between attributes and concepts. The test asks users to quickly and accurately categorize words or images and measures reaction time, such that faster responses indicate stronger associations than slower responses, suggesting how implicit attitudes can influence people's cognitive processes and behaviors. Diversity training [7] has been used to address implicit biases in organizational and educational settings (e.g., to improve attitudes toward women in STEM[31]).

Implicit biases have also been studied in medical school admissions. Capers et al. [12] measured implicit racial preference in medical school admissions committees using the black–white implicit association test [24] on 140 members of the admissions committee composed of faculty and students. The results show that all groups (men, women, students, faculty) display significant levels of implicit white preference. Education and training on implicit biases are suggested by the authors to reduce such biases. However, since these biases are unconscious, informing individuals about the existence of implicit biases has apparently limited impact during the decision making process.

Recent work also observed anchoring bias [32] in sequential evaluation of university applications. Echterhof et al.[18] demonstrated that individual reviewers can be anchored by their own recently made decisions such that a borderline application might be admitted if the reviewer just reviewed many under-qualified applicants before evaluating this application, and on the other hand, the same or similar application can be rejected if the reviewer just reviewed many qualified applications.

We hypothesize that a *real-time* intervention is a promising solution, by creating heightened reviewer awareness of their process and promoting reflection in the time and space of the analysis and decision. To the best of our knowledge, there has been no prior work that utilizes reviewers' real-time interactions with application packets to promote process awareness.

2.2 Analytic Provenance

Analyzing user interactions with visual analytic systems is a form of analytic provenance [43] which focuses on understanding users' reasoning process beyond analytic outcomes. The analysis of user interactions and provenance data has been used for various purposes [61], including evaluation of visualization systems, creating adaptive systems, model steering, replication of analysis sessions, report generation, and most relevant to our work, understanding the user. For example, studies have shown that user interaction is a powerful form for learning about users including their analysis strategies [17], personality traits[11], and subjective preferences on data[57].

Our work is similar to these prior works in that we also aim to obtain high level information about the user's analysis process from low level interaction such as mouse clicks. However, different from prior work in which the provenance information is used by the underlying system or researchers other than the user themselves, our work displays the provenance information to the user themselves to help reviewers maintain self-awareness of their decision making process.

2.3 Collaborative Visualization

The admissions review process represents a collaborative task where each committee member reviews parts of applications individually followed by a group discussion to aggregate individual ratings. Visualization techniques have been utilized to support collaborative data analysis [9, 25, 29] and group decision making[4, 27, 36]. Collaborative visualization can take place in many scenarios delineated according to space (co-located or distributed) and time (synchronous or asynchronous)[25]. Numerous systems have been designed to support collaborative visualizations in these different contexts. A common setting is distributed collaborative visualization where multiple users share visualizations remotely for solving problems as a team [3, 5, 33] or for social data exploration[26, 52]. In contrast to distributed systems, co-located collaborative visualization systems support face-to-face collaboration around a large shared display [30, 49] or across a set of mobile devices [34].

Recent works have designed visualization systems to support group decision making (e.g., [4, 27, 36]). A notable example by Liu et al[36] introduces a visualization tool that facilitates consensus building by displaying all group members' opinions across multiple criteria individually and in aggregate to help users identify points of disagreement. The above-mentioned techniques provide a useful starting point for designing and developing a collaborative visual analytic tool for admissions review that supports aggregation of reviewers' ratings on applications and helps consensus building on admissions decisions.

2.4 Bias in Visualization

Recent work to mitigate biases in the context of visual data analysis are perhaps the closest recent efforts to the present work. Our approach, however, is removed from eradication of specific biases and instead focuses on boosting individuals' awareness of the characteristics of their analytic process. Nonetheless, efforts toward addressing bias in visualization inform our work by providing contextual motivation and alternative approaches.

Bias has been actively studied in the visualization community recently, formalizing the types of bias relevant to visualization and visual analytics [16, 56]. Some efforts have examined the presence of particular types of bias in decision making processes with visualizations such as the attraction effect[15], priming and anchoring bias [13, 51, 54, 60], and confirmation bias [39]. Other recent work proposed computational metrics that can be applied to user interactions with data to quantify bias in real-time [19, 21, 55]. In the context of the application review process, prior metrics (e.g., [55]) fall short in that they require definition of an ideal baseline against which individuals' interactions can be compared. However, in the context of admissions, it is a noisy process which would be ethically difficult or impossible to define desired review behaviors. Hence, we choose a simplified approach based on surfacing measures of *focus*: the duration of interaction with different applications and different application components, without assigning a value judgment to the outcomes.

Apart from bias detection, researchers have also recently investigated methods to mitigate bias [8, 35, 59] by altering the framing of the task [14], or communicating bias metrics visually in real-time to increase the awareness of bias [40, 58]. The design of our system is inspired by recent work [40, 58] that captures and visualizes users' interaction history with data in real-time to promote reflection of one's data analysis process. While these efforts had mixed results in laboratory experiments, we posit that there is high potential for real-world impact in the context of graduate admissions.

Most relevant to the domain of this work, visualization researchers have identified potential biases (confirmation bias[41], the Halo Effect[42], and the avoidance of cognitive dissonance[2] to name a few) in the holistic review process in undergraduate admissions. Visualization strategies, e.g., presenting alternative visual representations of application attributes and applying single-text visualization methods on letters of recommendation and students' essays to identify salient and effective points are proposed to mitigate these possible biases[45, 46]. However, the proposed strategies are broad without empirical evidence of the effectiveness of specific strategies and have not been deployed in real admissions processes.

3 FORMATIVE DESIGN

The design and development of VIS4GRAD followed a user-centered approach [1] that involved close collaboration with two admissions committee chairs in the Computer Science Department of a private university. The department was motivated to improve equity and consistency in the admissions process. Our final system development resulted from multiple iterations, during which we interviewed the admissions committee chairs to learn the admissions process and identify requirements, design a prototype, gather feedback, refine the prototype, and implement the system.

3.1 Methodology

We first conducted **semi-structured interviews** with two admissions committee chairs in the Computer Science department to learn more about their graduate admissions decision-making process. The interview covered topics around the application format, data access, decision making criteria, and collaborative mechanisms, among others, with follow-up questions to dig deeper based on participants' answers. Each session lasted approximately 30 minutes.

Based on our understanding of the program needs, we next sketched a possible visualization solution and built a **preliminary prototype** from the sketches to ground the discussion on the program's goals for a review system. The first version of the interface is shown in Supplemental Materials.

In a subsequent session, we provided a **demonstration of the preliminary interface design**, followed by a continued semi-structured discussion on the interface. A survey questionnaire followed afterwards, including questions about the usefulness of each component of the system, comments on the features in the interface, and the willingness to use such a system in the future. Additional interviews were conducted with one committee chair after two additional iterations of design to collect **ongoing feedback on subsequent iterations** of the system. This process resulted in **characterization of the existing review process** (Section 3.2), **needs for process awareness** (Section 3.3) and **design goals** for the system (Section 3.4), described next.

3.2 Existing Review Process

The formative findings from semi-structured interviews illuminated the **existing review process**, which can be summarised as follows. One committee chair maps applicants to their interested faculty members stated in their application forms or to faculty members whose research area matches applicants' area of interest. Each applicant is reviewed by at least three admissions committee members in the initial review. The applicants' portfolios are concatenated into a single PDF file (one per applicant), and admissions committee members scroll through the file to review applications assigned to them.

The committee typically agrees beforehand about a set of criteria along which reviewers will rate candidates such as Academic Preparedness (performance in coursework that reflect readiness for success in graduate course and related academic processes), Research Preparedness (well-articulated aspirations and demonstrated ability or potential to conduct advanced, high quality research), Teaching Preparedness (relevant preparedness sufficient to be a teaching assistant for fundamental Computer Science courses), and Communication Proficiency (ability to communicate clearly in English verbally and written). These criteria are evaluated on a 0-5 scale where 0 indicates clearly fails the requirement and 5 indicates clearly exceeds the requirement. Reviewers enter their comments and ratings for each applicant in a shared spreadsheet. After the initial review, applicants who were "above the bar" are interviewed by at least one faculty member. The committee then meets to discuss and decide which applicants will be admitted, waitlisted, or rejected, referencing the spreadsheet of reviewer scores to anchor the discussion.

3.3 Process Awareness Needs

Based the interviews, we identified reviewers' *individual* and *group* needs for assessing (1) internal consistency in time spent across applicants and application components, (2) internal consistency in ratings/decisions across applicants, stratified by sensitive attributes such as race and gender. Inconsistency in time spent across applicants (1), for instance, could be not spending enough time on a certain application, neglecting a certain application component (e.g., recommendation letters), systematically spending more/less time on a certain group of applications across gender or race, and so on. While time spent alone is a noisy metric, it can nonetheless provide reviewers some point of reference to spark reflection. The system should also encourage reviewers to reflect on the decision outcomes and check internal consistency in their ratings (2). In this case, the review committee defines consistency in ratings to mean that applications with similar characteristics should be rated similarly. Inconsistency in ratings could be that applications with the same ratings in evaluation dimensions (e.g., research preparedness, communication, etc) received different overall ratings (e.g., competitive v. not competitive), or systematically rating a certain group of application as more/less competitive across gender or race, etc. In more extreme cases, this may manifest as an unbalanced racial and gender distribution of applicants recommended as competitive and not competitive (ultimately leading to unfair distributions of admitted and rejected applicants).

3.4 Design Goals

Based on our formative design activities, we derived the following **design goals** to support the department's desired framework for admissions reviews and the department's needs for increasing process awareness.

DG1. Facilitate independent review of applications. The system should enable admissions committee members to rate applications holistically and across a set of pre-defined dimensions individually.

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

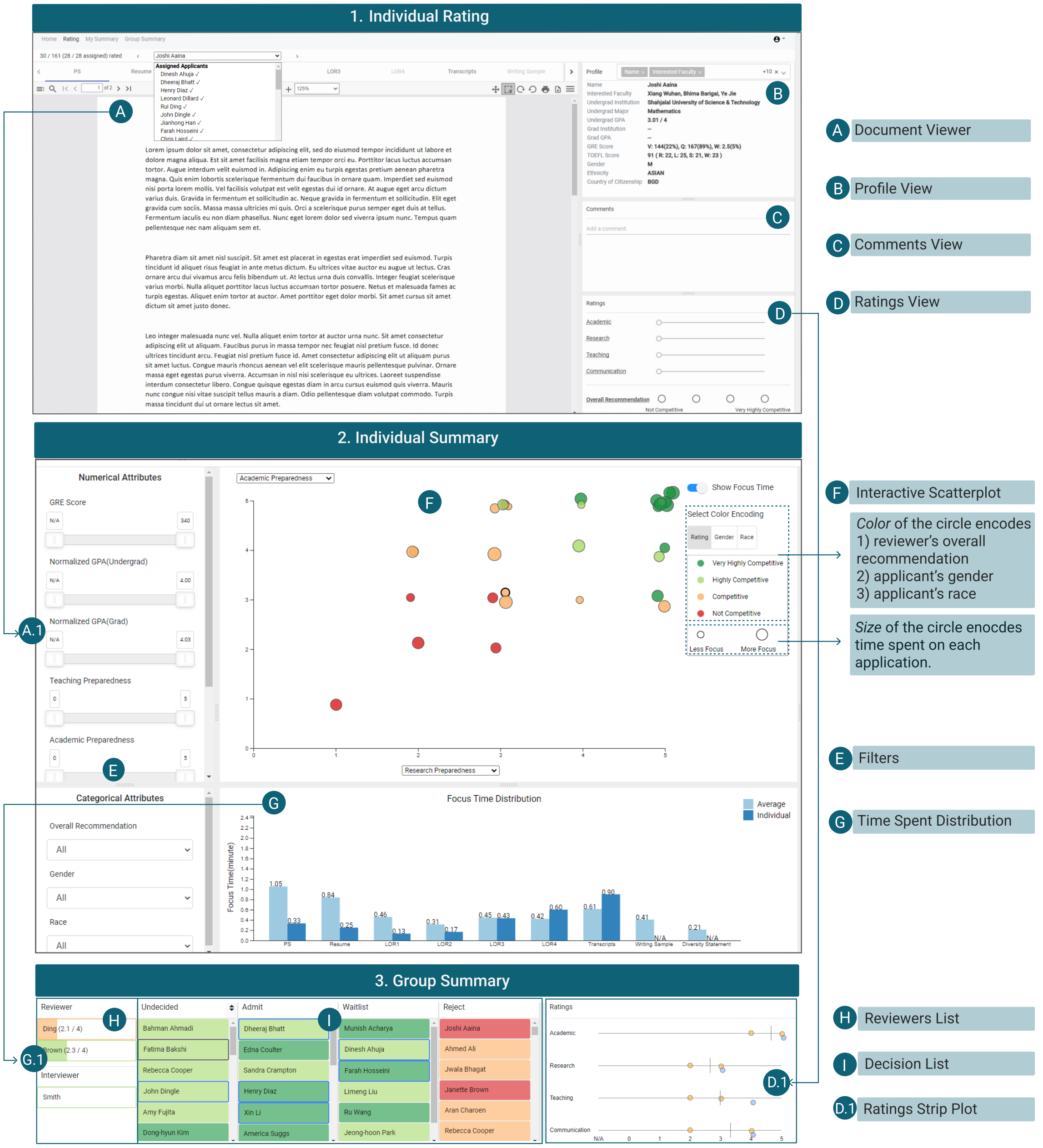
DG3. Facilitate group decision making. The system should facilitate group discussion of the applications during committee meetings by providing collated independent evaluations (ratings and comments) from individual reviews, and allow the committee to make ultimate admissions decisions for the applications, i.e., admit, waitlist, or reject.

DG4. Support group assessment of procedural consistency. In addition to supporting individual assessment of review behavior, the system should enable the department to assess the admissions review process in aggregate to facilitate assessment of admissions decisions in terms of adherence to departmental goals for diversity and equity.

DG5. Minimize the barrier to entry. The system should be visually simple and intuitive to increase adoption of the system over the status quo method for reviewing, and to ensure the system is usable by faculty members beyond the visualization domain.

4 SYSTEM

Based on our design goals, we developed a system VIS4GRAD (Figure 1) consisting of four separate tabbed pages, including a Home Page which shows basic information about the applicant pool and the reviewer's progress on assigned reviews; an Individual Rating Page where reviewers read and rate applications independently (DG1); an Individual Summary Page where reviewers can see a summary of their process as shown from their interactions with applicants (DG2); and a Group Summary Page which becomes available once all reviewers complete their independent reviews to provide the group an overview of their collective processes and facilitate finalizing admissions decisions (DG3, DG4).



Joshi Aaina

Assigned Applicants
Dinesh Ahuja ✓
Dheeraj Bhatt ✓
Henry Diaz ✓
Leonard Dillard ✓
Rui Ding ✓
John Dingle ✓
Jianhong Han ✓
Farah Hosseini ✓
Chris Laird ✓

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Est sit amet facilisis magna etiam tempor orci eu. Porttitor lacus luctus accumsan tortor. Augue interdum velit euismod in. Adipiscing enim eu turpis egestas pretium aenean pharetra magna. Quis enim lobortis scelerisque fermentum dui faucibus in ornare quam. Imperdiet sed euismod nisi porta lorem mollis. Vel facilisis volutpat est velit egestas dui id ornare. At augue eget arcu dictum varius duis. Gravida in fermentum et sollicitudin ac. Neque gravida in fermentum et sollicitudin. Elit eget gravida cum sociis. Massa massa ultricies mi quis. Orci a scelerisque purus semper eget duis at tellus. Fermentum iaculis eu non diam phasellus. Nunc eget lorem dolor sed viverra ipsum nunc. Tempus quam pellentesque nec nam aliquam sem et.

Pharetra diam sit amet nisl suscipit. Sit amet est placerat in egestas erat imperdiet sed euismod. Turpis tincidunt id aliquet risus feugiat in ante metus dictum. Eu ultrices vitae auctor eu augue ut lectus. Cras ornare arcu dui vivamus arcu felis bibendum ut. At lectus urna duis convallis. Integer feugiat scelerisque varius morbi. Nulla aliquet porttitor lacus luctus accumsan tortor posuere. Netus et malesuada fames ac turpis egestas. Aliquet enim tortor at auctor. Amet porttitor eget dolor morbi. Sit amet cursus sit amet dictum sit amet justo donec.

Leo integer malesuada nunc vel. Nulla aliquet enim tortor at auctor urna nunc. Sit amet consectetur adipiscing elit ut aliquam. Faucibus purus in massa tempor nec feugiat nisl pretium fusce. Id donec ultrices tincidunt arcu. Feugiat nisl pretium fusce id. Amet consectetur adipiscing elit ut aliquam purus sit amet luctus. Congue mauris rhoncus aenean vel elit scelerisque mauris pellentesque pulvinar. Ornare massa eget egestas purus viverra. Accumsan in nisl nisi scelerisque eu ultrices. Laoreet suspendisse interdum consectetur libero. Congue quisque egestas diam in arcu cursus euismod quis viverra. Mauris nunc congue nisi vitae suscipit tellus mauris a diam. Odio pellentesque diam volutpat commodo. Turpis massa tincidunt dui ut ornare lectus sit amet.

Name	Joshi Aaina
Interested Faculty	Xiang Wuhan, Bhima Barigai, Ye Jie
Undergrad Institution	Shahjalal University of Science & Technology
Undergrad Major	Mathematics
Undergrad GPA	3.01 / 4
Grad Institution	--
Grad GPA	--
GRE Score	V: 144(22%), Q: 167(89%), W: 2.5(5%)
TOEFL Score	91 (R: 22, L: 25, S: 21, W: 23)
Gender	M
Ethnicity	ASIAN
Country of Citizenship	BGD

Reviewer	Undecided	Admit	Waitlist	Reject
Ding (2.1 / 4)	Bahman Ahmadi	Dheeraj Bhatt	Munish Acharya	Joshi Aaina
Brown (2.3 / 4)	Fatima Bakshi	Edna Coulter	Dinesh Ahuja	Ahmed Ali
Interviewer	Rebecca Cooper	Sandra Crampton	Farah Hosseini	Jwala Bhagat
Smith	John Dingle	Henry Diaz	Limeng Liu	Janette Brown
	Amy Fujita	Xin Li	Ru Wang	Aran Charoen
	Dong-hyun Kim	America Suggs	Jeong-hoon Park	Rebecca Cooper

Color of the circle encodes
1) reviewer’s overall recommendation
2) applicant’s gender
3) applicant’s race

Size of the circle enocdes time spent on each application.

Fig. 1. Vis4GRAD supports admissions review with three pages. 1. Individual Rating: (A) Documents Viewer shows different application documents, (B) Profile View shows a set of attributes and a drop-down list for selecting/deselecting visible attributes, (C) Comments View and (D) Ratings View. 2. Individual Summary maintains the right-hand-side of the interface for profile, comments, and ratings of individual applicants, and replaces Document Viewer with (E) Filters, (F) Interactive Scatterplot, and (G) Time Spent Time Distribution Chart. 3. Group Summary maintains the scatterplot, profile, comments, and ratings, but replaces Time Spent Distribution with (H) Reviewer List and (I) Decision Lists. Note that all the application material displayed here is fake data for demonstration purposes.

4.1 Individual Rating Page

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

(A) Document Viewer shows PDF documents including personal statement, resume, letters of recommendation, transcript, and so on. Files are organized into separate tabs such that only one file is visible at a time to support subsequent meta analysis of review process by document (DG2). **(B) Profile View** shows tabular attributes of the applicant such as GPA, degree, major, etc. A set of default attributes are shown initially, and users can select/deselect attributes to be shown from a drop-down list if they would personally like to make their decision process blind to some sensitive attributes. This design is motivated by the feedback we received from the formative design process, where committee chairs expressed the need for flexibility to mask attributes such as gender and race that might bias their decisions and only reveal this information on demand. **(C) Comments View** allows reviewers to leave comments about the respective application, rather than externally taking notes as in existing practices. Comments are stored in the database and are subsequently shown in the Group Summary Page (Figure 1, bottom). **(D) Ratings View** allows reviewers to (i) rate the applicant on a set of factors (which can be pre-defined by the committee based on their review criteria) on a 0-5 scale and (ii) rate the overall competitiveness of the applicant on a 1-4 scale (in this case, Not Competitive (1), Competitive (2), Highly competitive (3), Very Highly Competitive (4)). Composing the application documents, comments, and ratings in a single interface makes it simpler to evaluate applications and adopt the system (DG5).

Interaction Logs record reviewers' time-stamped interactions on each applicant and each component of the application (i.e., personal statement, resume, letters of recommendation, etc). Specifically, users' interactions with the interface such as mouse move, click and scrolling are recorded to derive a reviewer's time spent (DG2). High level events such as page visibility changes (e.g., if the reviewer navigates to a different application such as a web browser) are recorded in order to identify time periods that users are not focused on the interface. Furthermore, the system applies thresholds to filter out outlier time periods (that are too short or too long) in order to reduce noise in derived time spent. If two interactions are within a very short period (e.g., only a few milliseconds), it can be regarded as random or unintentional. On the other hand, if the user has not interacted with the system for a long period, it is possible that the user has been distracted from the task.

4.2 Individual Summary Page

While the Individual Rating Page is intended to serve as an interface for completing existing reviewing tasks, the Individual Summary Page is intended to provide increased awareness of individuals' review process (DG2). This page can be accessed by reviewers any time during the admissions review cycle. It maintains the **Profile, Comments** and **Ratings** Views (Figure 1 B, C, and D, respectively) from the Individual Rating Page, but replaces the **Document** (A) panel with a data visualization panel (A.1), as shown in the middle of Figure 1. The visualization panel consists of the following components.

(E) Filters provide controls for filtering data by numerical or categorical attributes. In addition to the Profile attributes (such as gender, race, test scores, etc), users can also filter by their assigned ratings (e.g., for teaching preparedness, communication, etc) and overall recommendation of applicants (e.g., to view only candidates they rated as Competitive). **(F) Interactive Scatterplot** visualizes applications that have been reviewed where the x- and y-axes can be assigned from a drop-down list to represent variables such as GRE score, GPA, reviewer's ratings and overall recommendation and so on. Hovering on a point (applicant) in the scatterplot populates the Profile, Comment, and Ratings Views, and the Time Spent Distribution (described below) with the applicant's data.

The scatterplot is designed based on the need for providing visualizations that allow reviewers to observe patterns and outliers in their time spent and ratings along different dimensions (DG2). We chose a scatterplot because of its effectiveness in identifying patterns/outliers and its simplicity to be perceived by a general audience (reviewers beyond the visualization domain)(DG5). In addition to the x- and y-axis encodings, the points can be encoded by (i) size to indicate total time spent on each application and (ii) color to represent the reviewer's overall recommendation of an applicant, applicant gender, or applicant race (as shown in Figure 1 F). Collectively, these encodings enable reviewers to scrutinize trends, distributions, and outliers in their ratings and overall recommendations by race, gender, etc. For example, the size encoding can help reviewers identify if there are any applicants who they spent significantly less time reviewing so they can subsequently revisit the application.

(G) Time Spent Distribution shows a grouped bar chart depicting (i) the reviewer's average time spent on different documents and (ii) the reviewer's distribution of time spent on application components for the hovered applicant in the scatterplot. This view is designed to help users gain insights about the time they spent across different application components (DG2), e.g., to identify instances where they spent relatively more/less time on resume relative to personal statement (which may be intentional or unintentional) and identify outliers at the individual applicant level such as little or no review of a certain file for an applicant.

4.3 Group Summary Page

The Group Summary Page is similar to the Individual Summary Page in that it centers around an interactive scatterplot; however, it represents aggregated information for applicants across reviewers to support group level analysis of the admissions review process (DG4). The **Time Spent Distribution Chart** (G) is replaced with a Reviewer Panel (H) and three Decision Lists (I) (as shown in the bottom of Figure 1) to facilitate group decision making (DG3). The details are described below.

(F) Interactive Scatterplot remains similar to the Individual Summary Page except the size encoding of time spent is based on averaged time spent across reviewers, and the color encoding for rating is based on the aggregated overall recommendation among all the reviewers that rated the given applicant. The circle stroke style is used to encode rating agreement among reviewers such that a dashed stroke indicates reviewers have different overall recommendations on an applicant. This is designed to help the committee identify disagreements among reviewers easily to inform the focus of discussions. A new attribute, Admission Decision, is added in the x- and y-axes options and the color encoding options which can help the department to assess the admissions decisions in terms of departmental goals for diversity and equity in the final decisions made. Clicking on a point (applicant) in the scatterplot populates the Profile, Comment, and Ratings Views to facilitate group discussion on the application.

(D1) Ratings View in the Group Summary Page is updated to a strip plot representing each reviewer's rating scores on different factors, along with vertical lines indicating the mean score among reviewers for each factor. The ratings scores from the portfolio review phase and the interview phase are presented in a different color (i.e., portfolio review scores are represented in orange, and interview scores are represented in light blue). This view allows the group to easily identify the agreement and discrepancy in the ratings for a application from different reviewers. Hovering on a point in the strip plot will highlight the corresponding reviewer in the Reviewer List (described next), and hovering on a reviewer will highlight the reviewer's score in the strip plot.

(H) Reviewer List displays information about reviewers who reviewed or interviewed a given applicant when an applicant is selected in the scatterplot. Each reviewer's information is presented in a data cell including the name of the reviewer followed by the reviewer's average overall recommendation (on a scale of 1-4) provided to their assigned applications to facilitate calibration of scores. The reviewer's overall recommendation on the current applicant is encoded as the color of the cell's stroke which shares the same color encoding as the scatterplot, and the reviewer's average time spent on different applications is encoded as the length of the colored bar in the cell. An example is shown in the bottom of Figure 1 where two reviewers' and one interviewer's information is displayed. The average time spent and score can help facilitate the group discussion and decision making by understanding the relative toughness of reviewers' ratings.

(I) Decision Lists show four bins (Undecided, Admit, Waitlist, Reject). Applicants are all placed in the Undecided list initially and can be moved to and rank ordered in the relevant bin to finalize admissions decisions. Applicants can be moved into the appropriate bin by drag and drop from the scatterplot or from another bin and can be similarly reordered within the bin by drag and drop. In addition to manually ordering the applicants, sort functions are also supported in the Undecided list to facilitate prioritizing the order of discussion of applicants among the committee. The decision lists share the same color encoding as the scatterplot (Figure 1 shows the case when the color encoding is set to Rating (Average Overall Recommendation)). This page is disabled for individual reviewers initially until all reviewers have finished their own review process to guarantee that each reviewer considers and rates the applications independently.

4.4 Preliminary Feedback

To gauge the effectiveness of VIS4GRAD, we first obtained preliminary feedback using a heuristic evaluation [53] from five visualization experts who also have experience in graduate admissions. The system received positive scores (μ = 5.89 / 7) on all of the components (Insight, Confidence, Essence, and Time (ICE-T) [53]) of the heuristic evaluation framework. Participants also provided qualitative feedback about the system including general impressions on the system and suggestions for improvements. The details of the evaluation are attached in Supplemental Materials.

5 CASE STUDY

We conducted a case study in the Computer Science department at a private university where the system was used for the department's Ph.D. admissions reviews over a time period of roughly one month. This case study allowed us to assess real-world efficacy of our visualization for process awareness. The study methodology is described below and the study results are presented in the next section (Section 6).

5.1 Participants

Participants were recruited through email. The admissions committee chair sent out an email to all admissions committee members with an introduction of VIS4GRAD, a description of how to use the system, and a tutorial video that demonstrated the features of the system. The committee members were recommended (but not required) to use VIS4GRAD; they could still elect to use prior existing methods (described in Section 3.2) for their review tasks. There were two committee chairs and 12 committee members in total. The two committee chairs and 11/12 committee members used the system in some capacity during the admissions process. We were able to subsequently interview 11 participants (two committee chairs and nine committee members) after the admissions process concluded. They had 0 to 14 (μ = 4) years of prior involvement in admissions. We refer to participants in the "Results" section (Section 6)

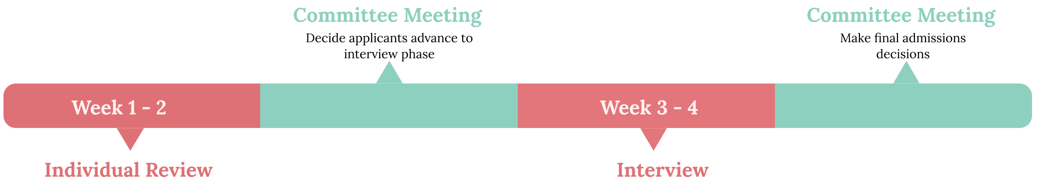


Fig. 2. Timeline for the admissions process.

as P1-13. We note that one of the authors was a member of the admissions committee. We include their data in our analysis in order to present a comprehensive view of the admission review process.

5.2 Dataset

There were 161 Ph.D. applications in total, and each application was reviewed by at least two committee members. Each application consists of general information including research interests, education background, test scores, etc.; personal statement; resume; up to four recommendation letters; transcripts; and optional files including writing samples and diversity statement. The applications were downloaded from the application portal and loaded to the database for Vis4GRAD by the first author before the system was available for reviewers.

5.3 Procedure

Participants were provided usernames and temporary passwords to securely access Vis4GRAD. The timeline of the admissions process is summarized in Figure 2. During the first two weeks of the admissions review phase, participants independently reviewed and rated applications assigned to them. After all committee members completed their review duties, the committee held a group meeting to discuss applications. Before the group meeting, the first author loaded the scores and notes provided in a spreadsheet from faculty members who did not use the system and enabled the Group Summary Page. During the meeting, the Group Summary Page was used to facilitate the discussion. The committee decided which candidates were not a good fit for the program (assigned to the Reject list) and the rest of the candidates (remaining in the Undecided list) were assigned to faculty members for an interview. The interviews took place in the subsequent two weeks. A second committee meeting took place after the interviews where the final admissions decisions were made.

After the admissions cycle concluded, emails were sent to all the committee members to invite them for an interview in order to gather user feedback. We interviewed two committee chairs and nine committee members who used the system. All the interviews were conducted via zoom and each interview lasted 30 to 60 minutes. The interviews started by showing a consent form to the participant and upon agreement, the interview was screen- and audio-recorded. After gathering background information, the interviewer asked the participant to login to the system and share their screen to facilitate a walk through of the system and discuss their experience (with different pages and features of the system) and provide suggestions for improvements. Following the interview, the participants were asked to complete a post-study questionnaire. The post-study questionnaire consisted of questions about the usability of the system described in Section 6.4.

5.4 Analysis and Coding

The first author transcribed the interview audio recordings, consulting the video recordings to resolve any ambiguities in the utterances. The research team used qualitative data analysis methods[37] to analyze the interview transcripts. Specifically, thematic analysis was conducted on the interviews through inductive coding [48]. Two authors independently coded two transcripts and discussed to develop a codebook. After refining coding definitions together, the the first author coded the remaining transcripts. The final codebook contains 36 codes in nine categories including System Usage, Review Strategy, Awareness and so on, included in Supplementary Materials.

6 RESULTS

We present quantitative analysis of participants' interaction logs and surveys and qualitative analysis of the post-study interviews. We organize results according to high-level themes of the case study, including increasing participants' **awareness** of their review process (Section 6.1), associated **behavioral** changes (Section 6.2) and changes in **decisions** (Section 6.3), consistent with [58]. In addition, system **usability** scores and feedback are presented in Section 6.4.

While 11 reviewers used the Individual Rating Page to complete review tasks, only five of them used the Individual Summary Page. Reviewers who did not actively use the Individual Summary Page during application review still provided feedback after interacting with the view during the interview. All of the reviewers interacted with the Group Summary Page directly or indirectly (through screen-share from the committee chair) during the committee meetings.

6.1 Awareness

We use the term **awareness** to refer to insights gained from reflecting on one's review process via interaction history on the applications.

6.1.1 Process Awareness. Participants found that features in the Individual Summary Page (specifically the distribution of time spent across documents in the bar chart and the interactive scatterplot) helped them to systematically reflect on internal consistency in their time spent across applications and application components (*"I was able to look at the amount of time that I spent over their different documents just to make sure I didn't miss anything."* - P9, *"I would also look at focus time and I would look for outliers."* - P12) and adjusted their review behaviors when they identified undesired behaviors. For instance, the focus time distribution chart helped P9 be **aware** that (*"I hadn't really spent time on their writing sample"*, led them to **behave differently** (*"So I went back and I had the chance to read over it"*, and made changes on some **decisions** (*"I think a student had uploaded something in the writing sample but they hadn't mentioned it on their CV, so it was a good chance to revise what I had scored for their research preparedness"*). The scatterplot helped P12 identify outliers in the time they spent on applications (*"I just didn't spend that much time on someone"*) and led to **behavior change** - *"I would try to go back and just look, just spend a little bit more time looking at them and see if I missed something."*

6.1.2 Outcome Awareness. In addition to assessing internal consistency of time spent, participants also used the Individual Summary Page to check internal consistency in their ratings (*"I check the different distributions just to make sure that I was sort of consistent in giving my overall ranking."* -P9, *"making sure that I'm internally consistent."* -P12). P9 tried to self-calibrate on the ratings (*"we had the chance to go back to our reviews, compare the ones that we had scored previously, and change our rating.*). P12 liked the color encoding of the scatterplot which *"was super helpful for helping me to calibrate how tough I am as a reviewer for the ratings."* The scatterplot also helped P12 identify outliers in the ratings (*"I identified outliers like someone that I rated as having high research preparedness but I did not rate them overall very highly."*) and revisited the applications (**behavior** change) – *"I would go back and look at their applications again"*. P2 thought that the scatterplot in the Group Summary Page is *"useful for giving an aggregate view, allowing you to find anomalies easily"* and is useful *"to see if the decisions are consistent"*. The committee planned to *"take a look at the scatterplot to evaluate our process overall"*.

6.1.3 Fairness/Bias. Participants found the system useful in terms of increasing awareness about procedural fairness. P5 liked that the profile view allows hiding attributes (*"I really like that... I basically turned off anything that I felt might bias my decision."*). P9 commented that the system made the review process *"fair on behalf of the applicants because we had the opportunity to compare different people, look at the demography and things like that."* P9 was interested in seeing *"was there really some sort of unintentional way of aspects that influence my decision."* By looking at the scatterplot in the Individual Summary Page with different combinations of X- and Y-axis attributes, the participant found that *"there's no bias towards any gender or race in my decision. That was good for me to to know."* P12 thought it was useful *"being able to sort of reflect on what biases might have come into play."* Furthermore, as described previously, these insights often led to changes in reviewer behavior and decisions.

During the interview, participants who did not actively use the Individual Summary Page during the admission process tried to interact with the interface and found the scatterplot *"is showing how I have reviewed people, some tendency of certain way to the other... if I have some gender bias or race bias."* (P4), could be used *"to make sure you're not admitting all men or something like that."* (P6), and the focus time could answer the question *"are you spending the right amount of time or at least enough time on all the different applicants?"* (P11), indicating the system has the potential to increase reviewers' awareness of bias during the review process, even if they did not ultimately use the system as such.

6.2 Interaction Analysis

In this section, we present our findings from quantitative analysis of user interactions with the system.

6.2.1 Interactions with Individual Summary. The system logged users' interactions with the Individual Summary Page, including hovering on the points on the scatterplot (Figure 1, F) (to see the applicant's information), clicking on the points (to revisit the Individual Rating Page), and clicking on the Overall Recommendation radio buttons (to modify overall recommendation). We analyze this interaction data to understand if and how users used this part of the interface, designed to promote reflection.

We found that reviewers visited the Individual Summary Page at different phases of the admission process, i.e., during the individual review phase and during the group meeting. Five reviewers visited the page, among which three actively interacted with the page during both phases and the other two reviewers interacted with the page only during the group meeting. Table 1 and Table 2 shows the number of distinct applicants reviewers hovered, revisited and made changes on overall recommendation (including the changes made after revisiting the Rating Page) in the two phases. Although the numbers were too small to make generalized statements, we further looked into what type of applicants reviewers tended to hover and click on. During the individual review phase, P9 hovered on almost all the applicants they reviewed, revisited applicants with low overall recommendation, and rated the applicants higher after revisits. P10 mostly hovered on applicants who they rated as Not Competitive and rated one of them higher afterwards. P12 hovered on and revisited more competitive applicants and both downgraded and upgraded some applicants. Based on the subsequent interview with this reviewer, this may be because they did not utilize the Not Competitive or Very Highly Competitive ratings much initially.

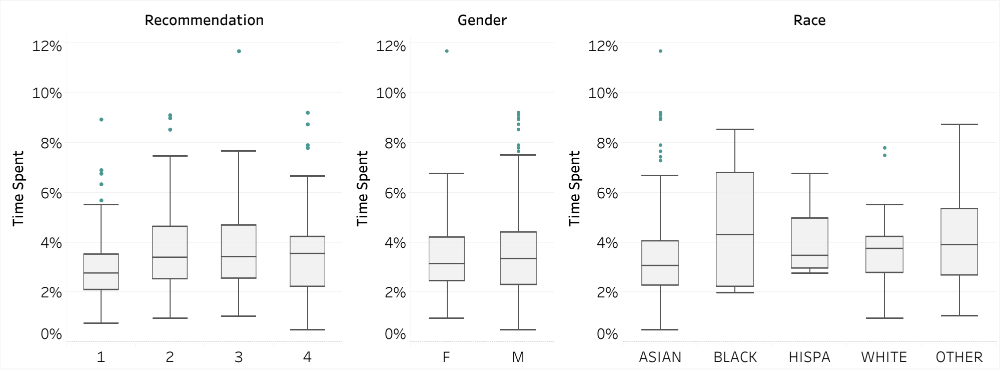


Fig. 3. Box plots comparing the time spent for applicants by different Overall Recommendation (left), Gender (middle), and Race groups across all committee members.

We also looked in to what color encodings and axes variables reviewers selected when interacting with the scatterplot. Most reviewers only used the default color encoding configuration (Overall Recommendation), except two reviewers also selected Race and Gender. Reviewers used different combinations of the x- and y-axis variables. The overall recommendation is frequently selected followed by research preparedness score, academic preparedness score and GPA.

Table 1. Participants' interactions with Individual Summary Page during the individual review phase.

	# Hover	# Revisit	# Changes in Recommendation
P9	27	3	3
P10	5	1	1
P12	13	3	2

Table 2. Participants' interactions with Individual Summary Page during the group meeting.

	# Hover	# Revisit	# Changes in Recommendation
P1	11	1	0
P5	5	0	0
P9	10	8	0
P10	22	4	0
P12	4	1	0

6.2.2 Time Spent. Figure 3 shows how reviewers (at the group level) spent time on applicants grouped by Overall Recommendation (left), gender (middle), and race (right). The time is normalized by each reviewer and is shown as a percentage (each reviewer's time spent on all applications they reviewed sums up to 100%). The race categories were created from the Ethnicity field from the application forms. The Other group included applicants with (1) multiple ethnicities specified and (2) applicants who did not specify their ethnicity. On average, reviewers spent less time on

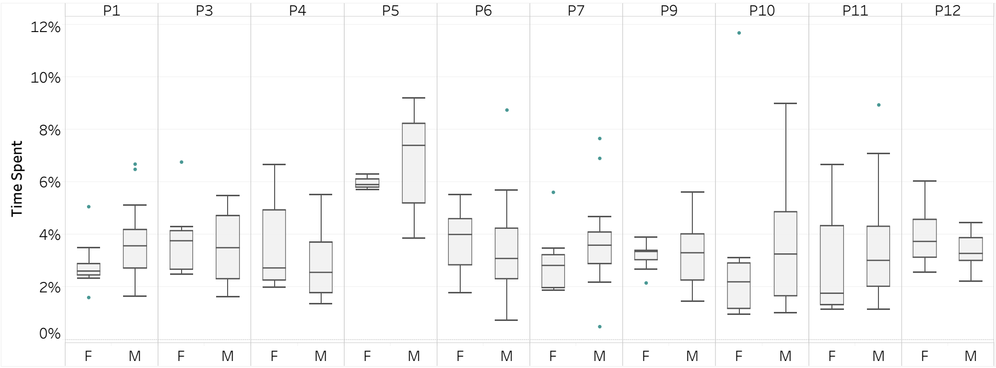


Fig. 4. Box plots comparing time spent on applicants grouped by reviewer and Gender.

applicants they rated as Not Competitive compared to applicants they rated as Competitive ($p = 0.024$) and Highly Competitive ($p = 0.017$). We found no significant difference in the time spent across gender, nor for time spent across most race groups, except that reviewers spent on average less time on Asian applicants compared to applicants in the Other and Black race categories ($p = 0.028$ and $p = 0.029$ respectively).

The aggregations at the group level, however, can dilute trends that may be observed at the individual level. Hence, upon further investigation, we found that in addition to the general trends identified at the group level, two of the reviewers also spent less time on applicants they rated as Not Competitive compared to applicants they rated as Very Highly Competitive. Although there is no trend observed at the group level on how reviewers spent time across gender, at the individual level, we observed vastly different trends among reviewers. As shown in Figure 4, some reviewers spent more time on Female applicants on average, while others spent more time on Male applicants. Only four reviewers spent more time on Female applicants on average. We note that although many comparisons did not result in statistically significant differences at the group level, the analyses of individual reviewer behavior can be cause for further scrutiny. We discuss this, along with potential explanations of these trends in Section 7.

6.2.3 Sequence Analysis. Motivated by recent work from Echterhof et al. [18] which observed anchoring bias in sequential decision tasks such as college admissions, we investigated how a reviewer's decision is impacted/anchored by previously made decisions in the graduate admissions review process. We collected 319 ratings made from 11 reviewers on 161 applications from the interaction logs. These individual ratings (overall recommendations) are dichotomized by treating ratings below or equal to a threshold as negative (0) and those above as positive (1). Following Echterhof et al.'s method [18], we calculate the Pearson correlation coefficient between the number of decisions made since the last positive decision and the current decision of the reviewer as a quantifiable measure of anchoring bias. The result on aggregated data from all reviews shows that there is a weak correlation ($r = -0.11$, $p < 0.05$) between the number of decisions made since the last positive decision and the current decision indicating reviewers can be biased by previously made decisions. At the individual level, we found that reviewers are impacted in different ways and at different magnitudes reflected by the differences in strength and direction of the correlation. Notably, we found that for three of the reviewers, their current decision is positively correlated with the number of consecutive negative decisions

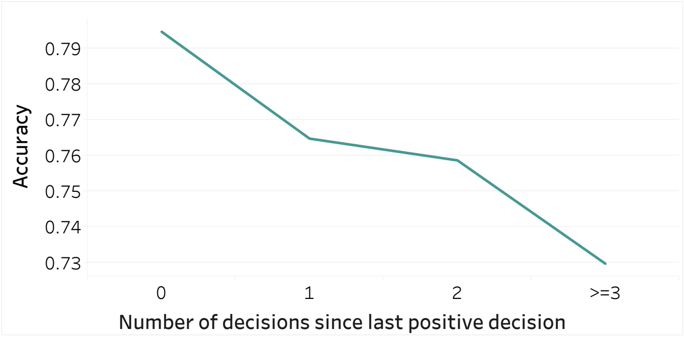


Fig. 5. The correlation between the number of decisions since last positive decision and the reviewers' decision accuracy. The accuracy (agreement between reviewers' decision and the group decision) decreases as the number of decisions since last positive decision was made increases.

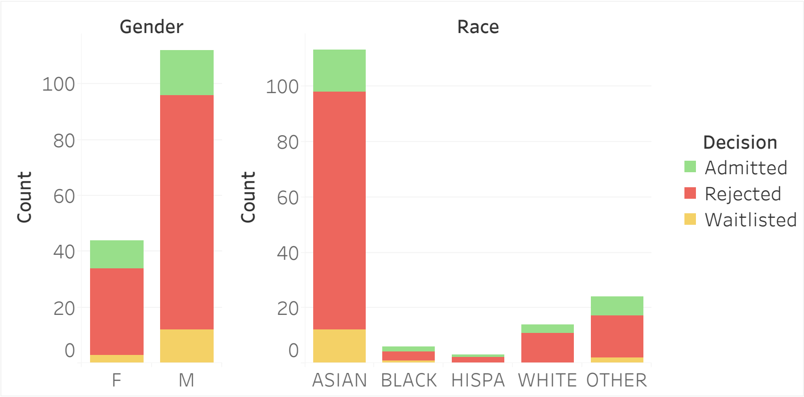


Fig. 6. The distribution of admissions decisions across gender and race.

made ($r = 0.20$, $r = 0.14$, and $r = 0.13$ respectively) indicating that when having reviewed many unqualified applicants the reviewer is more likely to rate the next applicant as positive.

We further investigated how such biases impact reviewers' decision accuracy, measured as the agreement between the reviewer's decision and the group decision. We found a negative correlation ($r = -0.97$, $p < 0.05$) between the number of decisions since the last positive decision and the accuracy of the current decision. As shown in Figure 5, when the number of decisions since the last positive decision increases, the decision accuracy decreases.

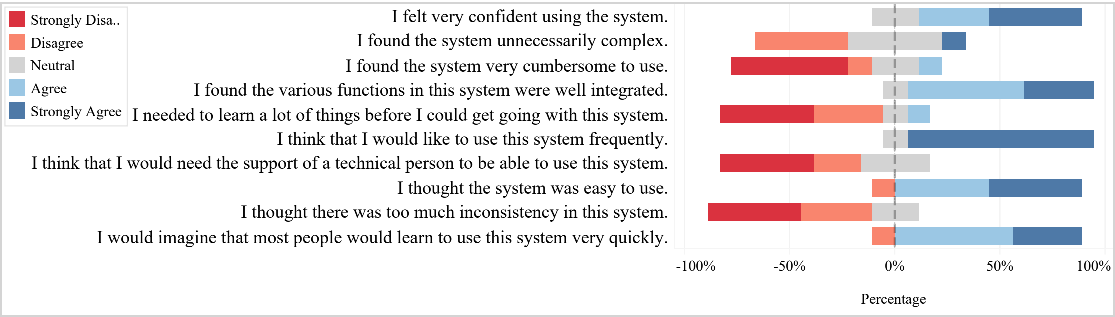


Fig. 7. Overall SUS scores of VIS4GRAD summarized in a diverging stacked bar chart.

6.3 Decisions

Figure 6 shows the distribution of the admissions decisions by gender and race. The distribution of admitted, waitlisted and rejected applicants by gender and race aligns to the underlying distribution of the candidate pool, i.e., there was no clear favor of a certain group when making decisions. The Chi-square test shows that the admission decisions are independent from gender ($p = 0.399$) and race ($p = 0.251$).

6.4 System Usability

In the post-study questionnaire, participants rated their overall experience (Figure 7) with the system as well as impressions on different features of the system (attached in Supplemental Materials.). The results of the questionnaire and the qualitative feedback from the interviews indicates that participants' overall experience with VIS4GRAD was positive.

Figure 7 summarizes participants' system usability scale(SUS) [10] scores. The average score is 78.3 indicating the system is "Good" (score above 68) [10]. Participants found the system made the review process faster and easier compared with the previous methods. P5 commented *"It's much easier than looking at the PDF"*, and P12 commented *"In general I found this is a lot faster and more pleasant way of going through the review process."* Participants also expressed interest to continue using the system in the future and suggested the system be used in other departments. For example, P2 said *"Overall, I think this was a great exercise and we would happily use it again in the future."*. P7 commented *"I wish other departments can use this system as well."*

7 DISCUSSION

Is Time Spent a Good Proxy for Bias? Time spent is a noisy proxy for bias, as noted by several participants. More time spent on an applicant does not necessarily reflect a negative bias. If a reviewer spends less time on a specific applicant, it could be due to unconscious bias, but it could also be due to other benign reasons. For instance, P11 noted *"I just know this candidate and I wrote her letter, and so I'm looking very little at her."* Other reviewers observed that *"the exceptions are the applicants who have previously reached out to us and we have already interviewed them and we already know them"* -P9 and *"just like in a conference reviewing setting, there are some manuscripts that are clear accepts and there's some manuscripts that are clear rejects."* -P11. Other factors also influenced time spent such as reviewer's familiarity with transcripts from foreign institutions (P6), general readability of other application components (P6), or varying lengths of documents like recommendation letters (P5). Reviewers tended to agree that *"most of the time is being spent on the murky middle"* -P11, which time spent in and of itself does not reflect.

There are a number of noisy factors influencing time spent as a proxy for bias. However, similar to the stated goals for Wall et al's bias metrics [55] and consistent with the goals of reflective design [44], our aim in Vis4GRAD is to promote individual and group *reflection* on potential biases. Thus while time spent is an imperfect proxy for bias, its representation in Vis4GRAD can cause reviewers to more carefully reflect on their review process. In the next section, we describe some possible ways to further increase engagement and reflection on bias.

Increasing Engagement and Reflections. As described in the case study (Section 6), only part of the committee used the features intended for reflection. Given our observations from the post-study interviews, we describe potential avenues for future efforts to increase engagement and reflection on review behavior. We observed that, although the formative system design occurred in close collaboration with committee chairs (Section 3), many committee members nonetheless found the Individual Summary and Group Summary pages to be visually overwhelming and chose not to engage (*"It looks scary to me ... I felt a little bit overwhelmed by what was going on"* -P11). Integrated explanatory features such as clickthrough tutorials, embedded videos, help pages or tooltips could increase the learnability for these pages. Additionally, future iterations of system design might utilize concepts such as progressive disclosure, presenting a minimal interface initially (with less critical views collapsed), and progressively add details and views on demand.

Furthermore, exploring the balance of mixed-initiative user interfaces [28] may be another promising direction. The system could gently nudge [47] participants to interact with reflective views or, more aggressively in cases of strict procedural goals, require it, e.g., using pop-up notifications that cannot be dismissed prior to engagement with the analysis.

Finally, committee chairs can pre-define bias analyses that are important (e.g., racial distribution of admitted applicants, time spent by applicant gender, etc.) and create default view configurations that individual committee members can use as a starting point to their exploration.

Ethics and Privacy. Although none of the review committee in our case study expressed privacy concerns, we must consider privacy with respect to appropriateness and individuals' willingness to share interaction data (time spent) which are shared with the committee in the Group Summary View. A reviewer's time spent on applicants is shared in two forms. First, the time spent on a particular applicant is shared in an aggregated form (i.e., averaged along with other reviewers who reviewed the applicant, where the average can be used to visually encode the size of circles in the scatterplot in the Group Summary View). Second, the average time spent across applicants they reviewed is represented in its disaggregated form as the length of the colored bar in the Reviewers Panel (Figure 1, H). Privacy concerns surrounding this form of data sharing can be informed through close collaboration and co-design with target users. In future work, we hope to provide flexible options for reviewers to opt in to the level of data sharing they are comfortable with.

Future Directions for Behavior and Decision Analysis. According to the analysis of reviewers' ratings on applications, we observe that reviewers generally rely more on part of the evaluation criteria (i.e., research preparedness and academic preparedness) when giving overall recommendations and in extreme cases, overly rely on a certain criterion like research preparedness. The inconsistency in reviewers' evaluation criteria can be undesirable when individual goals do not align with the department's overall goal. Analysis of the impact of each evaluation criteria for the overall recommendation can be shown to reviewers to allow assessment of whether the criteria they relied on are appropriate.

Future work can also seek to mitigate anchoring bias (described in Section 6.2.3) by tracking and analyzing reviewers' decisions. For instance, we can explore the use of visualizations of a reviewer's previously made decisions to increase the awareness of their anchor state. Machine learning techniques can be used to mitigate anchoring bias, as demonstrated in [18]: an algorithm that learns the anchor state of a reviewer and selects the next application to display in order to minimize anchoring effects increased the decision accuracy by 7%.

Metrics that can more accurately capture *unconscious biases* could lead to deeper reflections on an individual's process and potential biases. For instance, time spent on documents should be a function of factors like document length, complexity of vocabulary, formatting, etc. Thus future work can expand measures of bias beyond time spent on documents based on discrete interactions (e.g., mouse hovers and clicks) to can create a more accurate characterization by pairing discrete interaction metrics with passive attention metrics (e.g., using eye-tracking).

8 CONCLUSION

In this paper, we presented Vis4GRAD, a system designed to facilitate process aware admissions decision making. Designed alongside two graduate program admissions committee chairs, the system allows the admissions committee members to individually review applications, reflect on their review process, and collaboratively discuss, calibrate, and make admissions decisions. Reviewers' interactions with applications are recorded to capture time spent across applicants and application components and visualized to promote self-reflection of the review processes and increase awareness of potentially biased processes. We evaluated Vis4GRAD via a case study in the Computer Science department at a private university where the system was used for the department's 2022 graduate admissions cycle. While aggregate committee-level analyses suggested only minute, potentially noisy differences in review behaviors and decisions, individual review behaviors and decisions displayed more clear trends. We conclude that Vis4GRAD is a promising approach to increase awareness and affect changes in behaviors and decisions for individual admissions reviewers.

REFERENCES

[1] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.

[2] Elliot Aronson. 1969. The theory of cognitive dissonance: A current perspective. In *Advances in experimental social psychology*. Vol. 4. Elsevier, 1–34.

[3] Sriram Karthik Badam, Zehua Zeng, Emily Wall, Alex Endert, and Niklas Elmqvist. 2017. Supporting Team-First Visual Analytics through Group Activity Representations.. In *Graphics Interface*. 208–213.

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

[5] Aruna D Balakrishnan, Susan R Fussell, and Sara Kiesler. 2008. Do visualizations improve synchronous remote collaboration?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1227–1236.

[6] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[7] Katerina Bezrukova, Karen A Jehn, and Chester S Spell. 2012. Reviewing diversity training: Where we have been and where we should go. *Academy of Management Learning & Education* 11, 2 (2012), 207–227.

[8] David Borland, Jonathan Zhang, Smiti Kaul, and David Gotz. 2020. Selection-Bias-Corrected Visualization via Dynamic Reweighting. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1481–1491.

[9] Susan E Brennan, Klaus Mueller, Greg Zelinsky, IV Ramakrishnan, David S Warren, and Arie Kaufman. 2006. Toward a multi-analyst, collaborative framework for visual analytics. In *2006 IEEE Symposium On Visual Analytics Science And Technology*. IEEE, 129–136.

[10] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[11] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. 2014. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics* 20, 12 (2014), 1663–1672.

[12] 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.

[13] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 116–126.

[14] Evanthia Dimara, Gilles Bailly, Anastasia Bezerianos, and Steven Franconeri. 2018. Mitigating the attraction effect with visualizations. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 850–860.

[15] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2016. The attraction effect in information visualization. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 471–480.

[16] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2018. A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics* 26, 2 (2018), 1413–1432.

[17] Wenwen Dou, Dong Hyun Jeong, Felesia Stukes, William Ribarsky, Heather Richter Lipford, and Remco Chang. 2009. Recovering reasoning processes from user interactions. *IEEE computer graphics and applications* 29, 3 (2009), 52–61.

[18] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[19] Mi Feng, Evan Peck, and Lane Harrison. 2018. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 501–511.

[20] Gerd Gigerenzer. 2004. Fast and frugal heuristics: The tools of bounded rationality. *Blackwell handbook of judgment and decision making* 62 (2004), 88.

[21] David Gotz, Shun Sun, and Nan Cao. 2016. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 85–95.

[22] Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1 (1995), 4.

[23] Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California law review* 94, 4 (2006), 945–967.

[24] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[25] Jeffrey Heer and Maneesh Agrawala. 2008. Design considerations for collaborative visual analytics. *Information visualization* 7, 1 (2008), 49–62.

[26] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1029–1038.

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

[28] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[29] Petra Isenberg, Niklas Elmqvist, Jean Scholtz, Daniel Cernea, Kwan-Liu Ma, and Hans Hagen. 2011. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization* 10, 4 (2011), 310–326.

[30] Petra Isenberg, Danyel Fisher, Sharoda A Paul, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. 2011. Co-located collaborative visual analytics around a tabletop display. *IEEE Transactions on visualization and Computer Graphics* 18, 5 (2011), 689–702.

[31] Sarah M Jackson, Amy L Hillard, and Tamera R Schneider. 2014. Using implicit bias training to improve attitudes toward women in STEM. *Social Psychology of Education* 17, 3 (2014), 419–438.

[32] Karen E Jacowitz and Daniel Kahneman. 1995. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin* 21, 11 (1995), 1161–1166.

[33] Paul E Keel. 2006. Collaborative visual analytics: Inferring from the spatial organization and collaborative use of information. In *2006 IEEE Symposium On Visual Analytics Science And Technology*. IEEE, 137–144.

[34] Ricardo Langner, Tom Horak, and Raimund Dachselt. 2017. V is T iles: Coordinating and Combining Co-located Mobile Devices for Visual Data Exploration. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 626–636.

[35] Po-Ming Law and Rahul C Basole. 2018. Designing breadth-oriented data exploration for mitigating cognitive biases. In *Cognitive Biases in Visualizations*. Springer, 149–159.

[36] Weichen Liu, Sijia Xiao, Jacob T Browne, Ming Yang, and Steven P Dow. 2018. ConsensUs: Supporting multi-criteria group decisions by visualizing points of disagreement.

[37] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. 2018. *Qualitative data analysis: A methods sourcebook*. Sage publications.

[38] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences* 109, 41 (2012), 16474–16479.

[39] Atilla Alpay Nalcaci, Dilara Girgin, Semih Balki, Fatih Talay, Hasan Alp Boz, and Selim Balcisoy. 2019. Detection of Confirmation and Distinction Biases in Visual Analytics Systems.. In *TrustVis@ EuroVis*. 13–17.

[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.

[41] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[42] Richard E Nisbett and Timothy D Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology* 35, 4 (1977), 250.

[43] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. 2011. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 33–36.

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

[45] Poorna Talkad Sukumar and Ronald Metoyer. 2018. A visualization approach to addressing reviewer bias in holistic college admissions. In *Cognitive Biases in Visualizations*. Springer, 161–175.

[46] Poorna Talkad Sukumar, Ronald Metoyer, and Shuai He. 2018. Making a pecan pie: Understanding and supporting the holistic review process in admissions. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[47] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness.* Penguin.

[48] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[49] Matthew Tobiasz, Petra Isenberg, and Sheelagh Carpendale. 2009. Lark: Coordinating co-located collaboration with information visualization. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 1065–1072.

[50] Amos Tversky and Daniel Kahneman. 2013. Judgment under uncertainty: Heuristics and biases. In *HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part I*. World Scientific, 261–268.

[51] Andre Calero Valdez, Martina Ziefle, and Michael Sedlmair. 2017. Priming and anchoring effects in visualization. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 584–594.

[52] Fernanda B Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics* 13, 6 (2007), 1121–1128.

[53] Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. 2018. A heuristic approach to value-driven evaluation of visualizations. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 491–500.

[54] Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert. 2019. A formative study of interactive bias metrics in visual analytics using anchoring bias. In *IFIP Conference on Human-Computer Interaction*. Springer, 555–575.

[55] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.

[56] Emily Wall, Leslie M Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert. 2018. Four perspectives on human bias in visual analytics. In *Cognitive biases in visualizations*. Springer, 29–42.

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

[58] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2021. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 966–975.

[59] Emily Wall, John Stasko, and Alex Endert. 2019. Toward a design space for mitigating cognitive bias in vis. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 111–115.

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

[61] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. 2020. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 757–783.