



Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews

QIAN ZHU, The Hong Kong University of Science and Technology, China

LEO YU-HO LO, The Hong Kong University of Science and Technology, China

MENG XIA, Carnegie Mellon University, United States

ZIXIN CHEN, The Hong Kong University of Science and Technology, China

XIAOJUAN MA, The Hong Kong University of Science and Technology, China

People often take user ratings/reviews into consideration when shopping for products or services online. However, such user-generated data contains self-selection bias that could affect people's decisions and it is hard to resolve this issue completely by algorithms. In this work, we propose to raise people's awareness of the self-selection bias by making three types of information concerning user ratings/reviews transparent. We distill these three pieces of information, *i.e.*, *reviewers' experience, the extremity of emotion, and reported aspect(s)*, from the definition of self-selection bias and exploration of related literature. We further conduct an online survey to assess people's perceptions of the usefulness of such information and identify the exact facets (e.g., negative emotion) people care about in their decision process. Then, we propose a visual design to make such details behind user reviews transparent and integrate the design into an experimental website for evaluation. The results of a between-subjects study demonstrate that our bias-aware design significantly increases people's awareness of bias and their satisfaction with decision-making. We further offer a series of design implications for improving information transparency and awareness of bias in user-generated content.

CCS Concepts: • Human-centered computing → Empirical studies in HCI.

Additional Key Words and Phrases: bias-aware design, raising awareness, bias in user-generated data

ACM Reference Format:

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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 496 (November 2022), 31 pages. <https://doi.org/10.1145/3555597>

1 INTRODUCTION

People increasingly rely on online word of mouth (WOM) to learn about the quality of products or services when making decisions on the Internet [29]. Recent research showed that 87% of consumers read online reviews and 79% trust them as much as in-person recommendations [69]. However, various kinds of bias are embedded in user-generated contents that could affect viewers' decisions [7, 35]. Among them, the self-selection bias widely exists in user ratings and reviews [12, 48], and it is mainly caused by people's subjective participation in rating or writing reviews online [12, 48]. The typical reason for self-selection bias is people's tendency to give feedback only when they are

Authors' addresses: Qian Zhu, The Hong Kong University of Science and Technology, Hong Kong, China, qian.zhu@connect.ust.hk; Leo Yu-Ho Lo, The Hong Kong University of Science and Technology, Hong Kong, China, leoyuholo@gmail.com; Meng Xia, Carnegie Mellon University, Pittsburgh, United States, iris.xia@connect.ust.hk; Zixin Chen, The Hong Kong University of Science and Technology, Hong Kong, China, zchendf@connect.ust.hk; Xiaojuan Ma, The Hong Kong University of Science and Technology, Hong Kong, China, mxj@cse.ust.hk.

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

© 2022 Association for Computing Machinery.

2573-0142/2022/11-ART496 \$15.00

<https://doi.org/10.1145/3555597>

extremely satisfied or unsatisfied with the product or the service they had received [12, 48]. As a result, online rating/review becomes less representative [55] with the biased distribution of user feedback and people who are not aware of this bias may consequently be misled [8]. Hence, it is important to raise the awareness of self-selection bias and reduce its negative impact on users' decisions.

Some businesses try to mitigate the self-selection bias by sending emails to encourage people to leave comments, such as Yelp¹ and Tripadvisor². Such solicited reviews could provide a more representative set of user feedback [55]. However, the companies themselves also generated biased reviews through the creation of promotional user responses [64]. Another strategy adopted by companies is to ask customers to write about both the positives and negatives of their experiences or rate independently on multiple aspects of a product [70, 94]. Although these methods can provide comprehensive information for users, they mostly rely on financial or social incentives and thus may not scale well [55]. Moreover, these solutions are proposed from the perspective of business management, which aims to control the quality of users' feedback and improve the company's reputation. They do not take the end-users' best interest into consideration, who are at the end making decisions based on the biased data.

From a technical perspective, some works reduce the bias in user-generated data by statistical removing or changing the underrepresented samples [18], and other works try to mitigate the self-selection bias in user ratings/reviews by modeling, detecting and reporting suspected biases [101, 103]. These approaches consider bias as a kind of feature in data, but thus far, they can not capture it perfectly [104]. In addition, these technical approaches are often opaque for end-users who cannot understand the output of algorithms, thus causing some distrust issues [78].

The above two kinds of methods mainly focus on enhancing the sampling methods of user feedback or manipulating the data to reduce biases, but they can be difficult and costly to implement in practice. Furthermore, they can not eliminate the bias completely, which has been the nature of user-generated data. In addition, these two methods did not consider the perspectives of end-users who refer to the ratings/reviews to make their own decisions. For example, in the view of end-users, both strategies used by the enterprise and automatic algorithms at the back-end are invisible. Users might not realize the remaining bias behind the data processed by these methods when viewing the statistical mean ratings to assess the quality of an online product or service. Instead of trying to reduce the bias in data from the back-end, in this work, we proposed to make the potential bias "visible" to end-users to help them make informed decisions. We aim to **raise people's awareness of the self-selection bias** embedded in user-generated data – an approach to directly mitigating the impact of bias on people's decision-making [7].

To achieve this goal, we first propose to raise consumers' awareness of the self-selection bias in user ratings/reviews by making three types of information transparent, which are (1) *the reviewers' experience*, (2) *the extremity of emotion*, and (3) *the reported aspects* in user reviews. We distilled these pieces of information according to the literature and the definition of self-selection bias [6, 12, 41, 59]. Next, we conduct a large-scale survey ($n = 206$) to assess people's perceptions of these three types of information and identify the exact facets that are critical for their decision-making under the hotel booking scenario. Then, we design a visual display of these information aspects underneath user ratings/reviews and refine the design based on the feedback from a pilot study with two visualization experts and 12 users. We integrate the design into a prototype system that simulates typical online hotel booking experiences. After that, we explore how the bias-aware design may affect users' awareness of the bias and their decision-making process through a between-subjects

¹<https://www.yelp.com>

²<https://www.tripadvisor.com/>

study. Experimental results show that the design can raise people's awareness of the self-selection bias while also being thought of as helpful for making informed decisions. We discuss our findings and offer design implications at the end of the paper.

This work has three key contributions. (1) We identify three kinds of information related to the self-selection bias in online ratings/reviews and collect the specific information, including the reviewers' experience, emotion and aspects, that are needed to be visible to users. (2) We propose a bias-aware design and integrate it into an experimental platform with real-world data. (3) We derive empirical insights from a between-subjects study and offer implications for future designs to improve transparency and awareness of bias in user-generated content.

2 RELATED WORK

Our work contributes to the existing literature on how to support decision making based on online word of mouth (WOM). We describe the relevant literature from three aspects: (1) large-scale review analysis for decision-making, (2) reducing bias in data for decision-making, and (3) transparency for raising awareness and informed decisions.

2.1 User Review Analysis for Decision-Making

It has become common for people to make decisions based on others' opinions posted online, such as buying a product or booking a hotel online. However, it takes a lot of time for people to dig through large numbers of online reviews [29]. Existing works in the natural language processing (NLP) domain proposed automatic algorithms to extract representative information and/or general user attitudes from reviews, such as generating summaries of reviews [3, 87, 89] and classifying emotions in reviews by sentiment analysis [13, 16, 45]. Recent studies tried to define the features of helpful reviews [33, 97] by mining the factors related to user satisfaction [58] or analyzing the helpfulness by combining reviews and quantitative ratings jointly [21].

Based on these automated approaches, another body of work offered a visual interface or an interactive system for users to dig through online reviews in a more detailed manner [15, 22, 96]. They tend to present user reviews from multiple semantic perspectives and assist users in analyzing reviews from a higher level aspect. For example, Chang et al. visualized aspect-level data in hotel reviews for users to gain insights and analyze different types of customers [20]. ExtremeReader generates abstract visual summaries by offering a high-level structure of opinions in user reviews [93]. Considering the subjective aspects of user reviews, recent work designed an interactive visualization system with multiple views for domain experts to mine people's opinions [100]. These works provide users more space to explore and customize insights from user reviews. However, most of these systems or interfaces are complex for ordinary users and they are often designed for experts to analyze large-scale data.

The above works mainly focus on how to efficiently extract and represent user reviews, while they do not give much consideration to the potential bias hidden behind them. Considering the pervasive subjectivity in online reviews [44, 60], there is a strong need to reduce the impact of the self-selection bias on people's decision-making and help them make informed decisions. In this work, we leverage automatic approaches to analyze and extract the critical aspects of user ratings and reviews, and integrate the output into a bias-aware design for raising people's awareness.

2.2 Understanding and Mitigating Bias in User-generated Data

Previous researchers have indicated that user ratings and reviews are biased, not only caused by the companies (i.e., the organization/person who can manipulate ratings or reviews), but also by people who gave feedback [5, 24, 55]. One typical type of bias caused by people is the **self-selection bias**, which leads to more extreme positive and negative feedback in user ratings/reviews. While there

are several other types of bias caused by people or platform interventions, like anchoring bias, social influence bias, or rating bias by algorithms in the online community [23, 65, 88], in this paper, we mainly focus on the self-selection bias in user ratings and reviews [12, 48, 49, 55, 85].

To improve the representativeness of user feedback, researchers in the field of business and marketing designed different strategies to mitigate biases in data [1, 61, 71, 95]. These strategies include (1) sending emails to a random selection of users and encouraging them to write reviews [6, 55], (2) offering a relative comprehensive framework for users to give feedback (e.g., commenting on the pros and cons of a subject separately) [61, 95], and (3) selectively displaying a representative user feedback online by manipulating the orders [35]. However, these approaches are primarily designed for businesses with the aim of maintaining the reputation of a platform. They paid little attention to the end-users who refer to user ratings or reviews to make decisions. Additionally, biases may also be brought by such strategies, as their processes may include financial or social incentives [55].

Some other researchers proposed technical methods to handle biases in user ratings and reviews [9, 19, 82, 90, 101, 103]. These works aim to detect [19, 90] or model [82, 101, 103] biases in online ratings or reviews to mitigate the effects caused by biases. For instance, a recent work by Zheng et al. [103] identified and modeled biased user ratings based on textual reviews using deep learning models. Riyaz and Kriti used the *Kalman filtering technique* to estimate the sequential bias in user reviews [82]. There are also plenty of works that focus on building or designing unbiased algorithms in recommendation systems [52, 79] or rating systems [14, 47], which tend to deal with biases in algorithms rather than data. Although these approaches are proven to be useful to some extent, they still suffer issues of scalability, and they cannot reduce the bias completely due to the diversity of bias as well as the difficulty of designing unbiased algorithms [104]. Furthermore, they keep biases detected by the back-end algorithms that are opaque to end-users, and thus may cause issues of trust and explainability [27].

The self-selection bias in data are difficult to eliminate completely by automatic algorithms or data management strategies because it is caused by humans' self-selected behaviors. Additionally, the previous approaches typically performed operations that were invisible to the end-users and thus cannot help them to realize the potential bias and thus directly aid their decision making process. Therefore, it is critical to make people informed of the potential biased data while making decisions with user rating and reviews. In this work, we aim to raise people's awareness of the self-selection bias by putting the decision in the hands of users and reduce the risk of users being influenced by biased data rather than directly addressing biases.

2.3 Transparency for Decision-Making and Raising Awareness

Transparency can be used as a mechanism to help users make informed decisions with systems, and it usually comes with two goals: showing how and what [78, 102]. For the “**how**” situation, transparency can be used to make the process of how data is processed and analyzed visible to users. For example, in some recommendation systems, it is easier for users to make sense of the underlying algorithms when the process recommendation is shown [39, 98]. Transparency can greatly improve users' perception of the credibility and usefulness of a system [92], and empower the decision-making processes [31, 36]. The “**what**” type of transparency deals with the outcome of a system, revealing hidden data (if any) or displaying the reasons for the output [102]. Researchers improved transparency by providing supplemental information, such as visualization or explanation, to help decipher the existing results. Typical examples are post-hoc explanations that help users to understand the information provided by algorithms [25, 74]. Facing with transparent information provide, users trust the system more and become more satisfied about their decisions [26]. In this

paper, we mainly focus on the second form of transparency (“what”) and explore how to use it to increase people’s awareness of the self-selection bias in user-generated data.

Transparent design of data plays a vital role in users’ decision-making in the area of visualization [11]. Similar designs for users to track and analyze data include the Google Dashboard [42] and Mozilla’s Lightbeam [67]. Similar works have explored ways to give users back control of their data through tools that enhance transparency, as summarized in [50]. There also exist many works that leverage visualization to provide users with transparency about particular aspects of the data in different scenarios [4, 40, 56]. For instance, Zavou et al. informed users via a “chord diagram” visualization that helps users understand how their data is treated by third-part cloud-hosted services [99].

In addition to enhancing data understanding and informing user decisions, transparency is also beneficial for **raising users’ awareness** of the hidden patterns behind data. A previous work pointed out a critical need for increased data transparency because the layperson might not be fully aware of how data aggregation was done by third-party services [77]. Some other works emphasized that users should be aware of how systems or algorithms process user data and called for higher transparency [35, 78]. Transparent features were also used to improve users’ awareness of their behaviors [86] and personal privacy issues [34, 54]. Visualization can be used to promote transparency and raise people’s awareness of biases for informed decisions. The latest work by Arpit et al. proposed to raise users’ awareness of their biased behaviors during data analysis for decision-making by showing the interaction traces [72].

In this paper, we set the goal of raising people’s awareness of the self-selection bias in user ratings and reviews for informed decisions. We propose a bias-aware design based on user ratings to reduce the impact of self-selection biases on people’s decisions.

3 BACKGROUND

In this section, we first introduce the definition and the threat of self-selection bias – the focus of this paper – in user ratings and reviews. Then, we propose to help raise people’s awareness of this type of bias when referring to user-generated content by making information transparent.

3.1 Self-selection Bias on the Web

In this work, we target the self-selection bias in user-generated data: **people with extreme experience (i.e., positive or negative experience) are more likely to give their feedback online than those who have a moderate experience** [12, 59]. Online ratings and reviews widely suffer from the self-selection bias [5, 7] and thus often fail to provide information about the quality of products/services that represents the general opinions of the entire user base [28, 59].

The self-selection bias is mainly caused by people’s self-selected behaviors [12, 59] that can be unconscious or triggered by people’s fast, instinctive processing system [37, 53]. For example, people could self-select to report or not report based on their experience (extreme or not extreme), and choose to merely report the aspects that leave the strongest impression. Such kind of self-selection behavior is part of human nature [30], thus, the resulting bias in user-generated data is hard to prevent and remove completely by algorithms or data management [46].

Furthermore, later users – who refer to the ratings/reviews – may find their personal experiences with the products/services inconsistent with their expectations established based on existing user reviews, as the biased feedback might not give a complete, up-to-date picture. The big expectation-experience disparity may cause a vicious circle of extreme feedback online, which hinders people’s decision-making process [6].

Hence, it is critical to reduce the chance of people being “*victims*” of the self-selection bias when referring to online ratings/reviews for their decisions. Raising people’s awareness of the bias

can guide us toward a solution to deal with biases in user-generated data [7, 30, 46]. In the next subsection, we introduce how we raise people's awareness of the bias by making them realize that certain information is over-represented in the sample of user reviews.

3.2 Raising Awareness of Bias by Making Information Transparent

We aimed to raise people's awareness of the self-selection bias by making three types of information that potentially reflect self-selection bias transparent in user ratings/reviews. We distilled three types of information by reviewing the literature [6, 7, 12, 41, 44, 55, 59, 97]: (I1) the distribution of **reviewers** with difference experience who choose to share their opinions [7, 41], (I2) the distribution of **emotion extremity** [6, 12, 55], and (I3) the distribution of **reported aspect(s)** in user reviews [44, 59]. We introduce these pieces of information and why they need to be presented to users in details below.

- We concerned the **composition of reviewers (feedback providers)** because those who leave comments online often cannot represent the entire population of people who have tried the products/services [41]. There exists the “*silent majority*” who select not to give ratings or reviews on the Internet [7]. Thus, we need to keep readers of online ratings/reviews informed of the characteristics of the reviewers.
- We cared about the **emotion extremity** of feedback because people tend to speak out when they have very strong feelings about something, which only constitutes a relatively small portion of the experiences of the majority [12, 59]. Thus, it is necessary to make users aware of the polarity and intensity of emotions behind the user ratings/reviews.
- We considered the **reported aspect(s)** of user reviews because people can self-select to give feedback based on their own preferences [44]. They can be more sensitive to or impressed by one or several specific aspects of a product/service and write reviews (or give ratings) accordingly rather than equally weighting all aspects [59]. Showing the categories of aspect(s) that formed the basis of current user ratings/reviews could facilitate new users weight the feedback based on their own needs and interests.

Transparency can be an efficient way to raise people's awareness of the biased information from the perspective of how users perceive user ratings and reviews [30]. Before considering how to make the information transparent to raise awareness of the bias, we need to know exactly (1) how people perceive the three types of information and (2) what are the specific facets they would like to view about these three kinds of information when making decisions online. We conducted a formative study to collect users' opinions for these questions to inform the bias-aware design.

4 FORMATIVE STUDY

The goal of the formative study is to explore how people perceive the three types of information, and what exact information they care about when referring user ratings and reviews. We selected the hotel booking scenario among various scenarios because, comparing with other scenarios (e.g., watching a movie), people make more careful decisions in booking hotels since they pay relatively more money and time to buy and experience the service [62].

4.1 Procedure

We conducted an online survey with 206 participants who had booked a hotel online in the past two years. We performed the survey via Prolific³, which is a crowd-sourcing platform designed specifically for online academic research. Prolific has been shown to provide better quality data from diverse participants than other platforms [76].

³<https://www.prolific.co/>

We firstly set up a screening session to find participants who have experience in booking hotels online in the last two years. There are 273 of 400 participants who passed the screening session and 219 of them chose to participate in the formal survey. We collected 206 valid responses after excluding the participants who finished the survey in an extremely short time (<2 minutes).

The content of the survey can be divided into three parts.

- (1) Participants first answered the questions about demographics and their habits of booking hotels, such as reasons of booking hotels, platforms they used, etc. We collected their habits as the guide for the later development of prototype systems with the bias-aware design.
- (2) Then, we collected participants' perceptions of the three types of information (**I1**, **I2**, **I3**) from two perspectives: how do they perceive the information and to what extent do they make decisions based on the information when booking a hotel. To do so, we provided statements for each type of information and let participants to rate their agreement or disagreement on a scale of 1 to 7 (from strongly disagree to strongly agree).

For example, the statements for the **extremity** of emotion (**I2**) are expressed like: “*The extreme positive ratings/reviews are accurate*”, “*The extremely positive ratings/reviews affect your likelihood of choosing a hotel*”, etc. The word “positive” can be replaced by “negative” or “moderate” for different statements. To get the comprehensive perceptions, we also provide statements using different words beyond “*accurate*”, such as “*trustworthy*”, “*reflect people's opinions*”, “*reflect the quality of hotels*” or “*reflect different aspects of hotels*”.

- (3) In the third part, we collected the specific information participants care about when booking hotels online. We set questions like “*How did you filter reviews when you book a hotel online?*”, “*What kind of information of the reviewers you would like to view?*”, etc.

All questions are compulsory for participants and they need to return a complete code of Prolific after finishing the survey.

4.2 Participants

We controlled the acceptance rate (>95%) of all participants in Prolific. Each participant was paid \$2 for the survey that took eight minutes on average ($MAX = 11$, $MIN = 6$). These participants are all US-based (81 females, 122 males, 3 preferred not to disclose) with a mean age of 34 years ($MAX = 19$, $MIN = 71$). They came from different areas of the US (West 29.61%, Northeast 23.79%, Southeast 22.33%, Midwest 15.53%, Southwest 8.74%). 52.43% of them were full-time employed, 17.96% were students, 10.19% were employed part-time, 6.80% were unemployed, 4.85% were self-employed, 2.43% were retired, and 5.34% selected “other”. Participants selected their highest educational level from the following options: high school degree (7.77%), the college without a degree (17.96%), associate degree (7.28%), bachelor's degree (41.26%), master's degree (20.39%), professional degree (2.43%), and doctorate (2.91%).

4.3 Findings

4.3.1 People pay attention to the background information of reviewers and mainly focus on reviewers' rating experience. We found 58.77% participants reported that they cared about the information of reviewers and 40.53% participants considered the information of reviewers that affects their decisions when reading user reviews online. Among various kinds of information about the reviewers (e.g., age, gender), 61.92% participants thought the rating experience of reviewers is critical for their decision-making online. The rating experiences include (1) the number of reviews written by a reviewer (31.07%), and (2) the number of “*helpful votes*” received by a reviewer (30.85%).

4.3.2 People's decisions are more likely to be affected by extreme positive or negative feedback than moderate ones. Participants thought the extreme positive ($MEAN = 5.67$, $SD = 1.16$) and negative

($MEAN = 5.5, SD = 1.22$) ratings and reviews affect their decisions more than the moderate ones ($MEAN = 4.99, SD = 1.42$). Thus, their decisions were influenced more by extreme positive or negative feedback. Also, participants thought the emotions expressed at different extreme levels in user ratings/reviews are trustworthy that can reflect the overall opinions of users, as well as the quality of hotels.

4.3.3 People care about the reported aspects when referring user reviews. In addition to the above findings, we found that participants preferred to filter reviews by their preferred aspects (23.39%) and would like to read user reviews that have a photo so that they can check the detailed aspects of a hotel (27.58%). People's preferred aspects of hotels are diverse (e.g., food or facilities), and they may associate the ratings they see with the aspects they care about, without being prompted for additional information [44].

In conclusion, we found people mainly care about the reviewers' rating experience, including the number of written reviews and votes obtained, extreme feedback with positive and negative emotions, and the specific aspects reported in user reviews, and these kinds of information indeed affect their decision-making when referring user ratings/reviews.

5 USER-CENTERED BIAS-AWARE DESIGN

To raise people's awareness of the self-selection bias in user ratings/reviews, we provided a bias-aware design based on existing rating bar charts. In this section, we first describe our design requirements and then offer three alternative designs accordingly. We present the final design we chose according to the insights obtained from the pilot study, as well as the details refined in the design based on users' feedback.

5.1 Design Requirements

DR1: Disclose the distribution of the information behind user ratings. The current design of user ratings online merely provides the average rating and the number of different ratings (e.g., number of 1-5 stars). Therefore, it is hard for people to realize the self-selection bias hidden under user ratings and reviews with the current design. According to Section 3 and Section 4, we aimed to disclose the distribution of reviewers' rating experience (I1), the extreme level of emotions (I2), and different reported aspects (I3) in user reviews to raise people's awareness of the self-selection bias. We discuss the details on how to extract these three types of information in Section 6.3. Making people informed of how these aspects are distributed in user reviews can reduce the influence of bias on their decisions.

DR2: Use simple and common visualizations for lay individuals to support their decisions. The target audience of our work is people who refer to user ratings/reviews online to make decisions, and they usually have no expertise in data visualization. Hence, the design should be simple and make the best use of common visual representations that people commonly encounter in their daily lives. Additionally, we considered the practicality and generality of the bias-aware design by discussing with two visualization experts, as it aims to support people's daily online decisions. Therefore, we proposed three design alternatives based on the conventional design of user ratings (i.e., bar chart) rather than creating an entirely new design [75].

DR3: Enable filtering functions with the transparent information for exploring user reviews. When making purchase decisions, people spend the majority of their time exploring user reviews to mine suitable opinions for their personal decisions [93]. According to the formative study, the information we aimed to disclose in the bias-aware design is also critical for people's decisions. Thus, providing users with filtering functions based on transparent information is necessary to facilitate efficient decision-making.

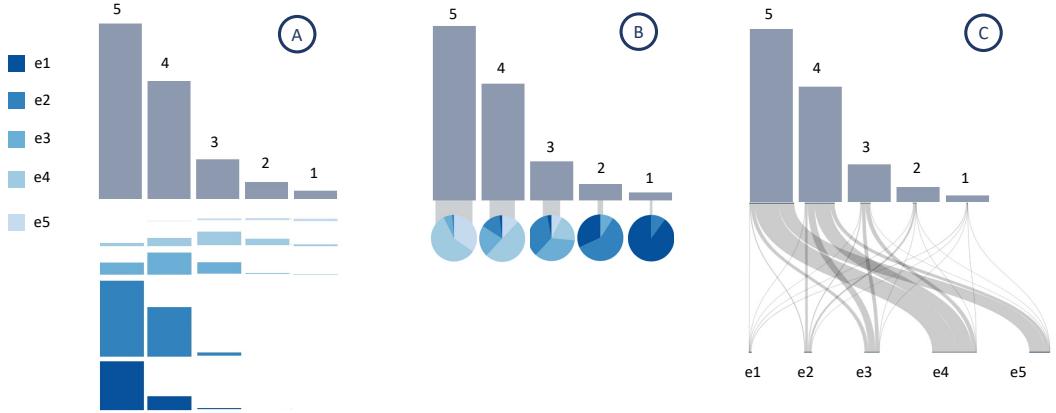


Fig. 1. The alternative designs for the user ratings, where the proportions of information are encoded by: (A) a stacked bar chart design, (B) a pie chart design, and (C) a Sankey chart design.

5.2 Pilot Study of Three Alternative Designs

5.2.1 Three Alternatives. To explore how to effectively show the distribution of information with user ratings (**DR1**), we proposed three alternatives (Fig. 1). We first collected various visualizations that display data distributions based on the literature [68], as well as the Internet⁴. Then, we discussed with two visualization experts for two rounds to select suitable visual representations that can be understood by ordinary users while conveying the information effectively (**DR2**). Finally, we selected the stacked bar chart (Fig. 1 (A)), pie charts (Fig. 1 (B)), and a Sankey chart (Fig. 1 (C)) as alternatives to show the distribution behind user ratings. We used the real-world data of a hotel to demonstrate the alternatives. Details of extracting the transparent information and calculating the distribution of different categories are illustrated in Section 6.3.

The three alternatives encode the same data derived from the user ratings and reviews of the example hotel. Since there are different types of transparency information to show (e.g., reviewer demographics or review sentiment), we use abstract variable labels (i.e., ■ e1 – □ e5 in the legend on the right of Fig. 1) to represent the different categories under each type of transparent information. For example, if the information currently encoded in the design alternatives is the emotion extremity (I2) of the user ratings as recognized by automatic algorithms, then ■ e1 denotes the category of extreme positive emotion while □ e5 corresponds to the extreme negative emotion category. Based on the legend, in Fig. 1 (A), the distribution of ■ e1 – □ e5 categories, regardless of data behind, is encoded by the length of stacked bars displayed under the user rating bars. The second alternative (Fig. 1 (B)) shows the distribution of emotion (I2) categories by a pie chart under each rating bar, and the third alternative in Fig. 1 (C) uses the Sankey chart, which encodes the distribution by the thickness of the branches.

5.2.2 Pilot Study. To evaluate the three alternatives and find the most appropriate design, we conducted a pilot study in a one-to-one interview manner with 12 participants (six females, average age = 24), recruited via university emails. They had different backgrounds, including computer science (6/12), landscape architecture (1/12), finance (1/12), graphic design (2/12) and online education (2/12).

⁴<https://datavizcatalogue.com/search/distribution.html>

We started the pilot study with a brief introduction to the background of our work and the bias-aware design with three types of information. Then we asked each participant to choose the most appropriate design and give reasons. We also asked participants how to interpret the additional information of each design by letting them imagine using these designs for booking a hotel online. We obtained participants' choices and feedback on each alternative during the study. One author analyzed the results independently and discussed them with the other three authors to identify user choices and needs.

5.2.3 Result. The second design in Fig. 1 (B) got the most votes (8 out of 12) among the alternatives, as it was simpler, space-saving and easier to understand than others reported by participants. We analyze the reasons of participants' selection in this part and we will introduce the details of the design (B) in next subsection 5.3.

During the study, we observed that people interpreted information in the designs from two aspects. One is the distribution of information (i.e., ■ e1 – □ e5) aligned with each rating bar, and the other is the distribution of one category (e.g., ■ e1) across different rating bars (i.e., 1-5 points). We called these two aspects “**vertical**” and “**horizontal**” information acquisition, as the former focuses on the information under one rating bar while the latter spotlights one category across all different ratings.

Participants gave their reasons for selection according to these two aspects. For example, although the stacked bar chart (Fig. 2 (A)) provides an intuitive way for users to compare different categories (i.e., e1-e5) across different ratings “*horizontally*”, it is hard for people to check the data “*vertically*” under one rating, especially for the 1-star bar. Participants (7 out of 12) said they focused more on the 1-star bar than the other ratings (2–5 stars) when they imagined using the design for online decision-making. However, the 1-star bar in Fig. 1 (A) is too thin for them to view and interact with. In addition, there are five participants reporting that the design in Fig. 1 (A) has poor scalability when the number of categories increases as it will take too much space to show the stacked bars under the user rating. For the alternative in Fig. 1 (C), nine participants reported that it has too many lines, which makes their interpretation process difficult. Even if the design clearly shows the proportion of different categories under a rating bar by the thickness of branches, the intersection of these lines makes it difficult for people to compare and analyze data “*vertically*” and “*horizontally*”. Moreover, four participants said that the alternative in Fig. 1 (C) is a bit complex for them as they had not used the Sankey chart previously (DR2).

Therefore, we added an additional forth design requirement **DR4** based on participant's feedback. **DR4: Support interactive data comparison among different categories (vertically) and user ratings (horizontally).** As users need to view and compare the transparent information with user rating bars from these two aspects, it is necessary to provide an efficient way for people to access the distribution *vertically* and *horizontally*. We further improved the design in Fig. 1 (B), which is selected by most participants, based on the **DR4**. We detail how we improve the design with interactive functions in the next subsection 5.3.

5.3 The Final Bias-aware Design and Interactive Functions

As shown in Fig. 2, we present the final design of the transparent information with user ratings. The design combines pie charts with user ratings that clearly show the distribution of information (e.g., ■ e1 – □ e5) *vertically* in user reviews behind each bar (**DR1**). It shows the categories across user ratings *horizontally*, leveraging the interactive design in Fig. 2(A3). The pie chart is a conventional chart showing distributed data, which is familiar to people in daily life (**DR2**). The legend on the right-hand side of the design shows the color scale associated with each category (i.e., ■ e1 – □ e5) of the information (i.e., the extremity of emotion). To coordinate the proportions of elements in the

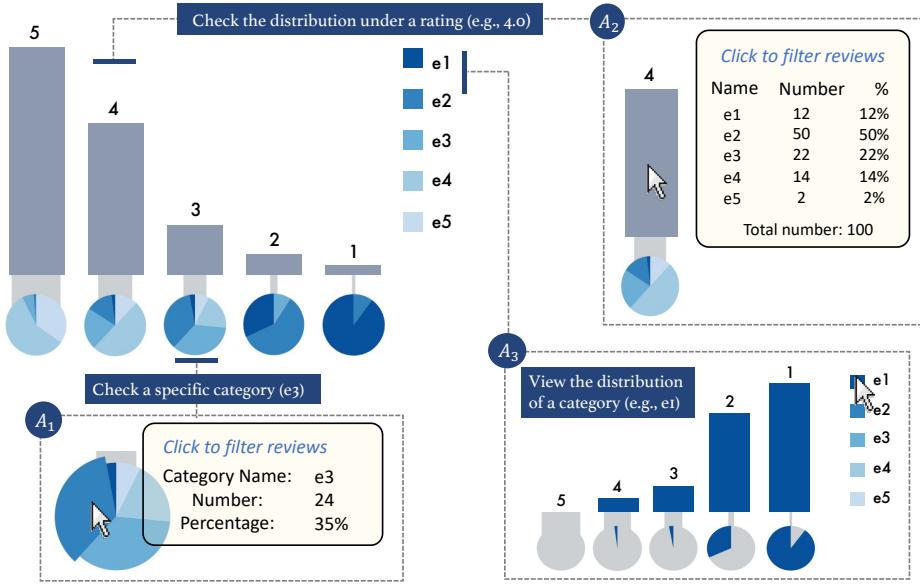


Fig. 2. The bias-aware design and with three interactive functions. The distributions are encoded in pie charts for each rating bar. (A1) shows the hover effect on the pie chart; (A2) displays the hover effect on a rating bar, and (A3) is the result triggered by hovering on the legend rectangles. The user rating bars in (A3) can vary according to the selected category (e.g., e1) by hovering over a category in the legend.

design and improve space utilization, we keep all pie charts the same size and use the thickness of the (light grey) lines linking pies and bars to represent the number of ratings/reviews for each pie.

We designed three interactive functions for the design in Fig. 2 according to the participants' requirements in the pilot study. These functions could help users compare the detailed information with the design (**DR4**) while offering filtering functions to filter user reviews in a specific aspect (**DR3**). In Fig. 2 (A1), we added a zoom effect on each sector of a pie chart so that users can easily hover over them to view the details of each category in a floating text box and filter user reviews by clicking a sector. Similarly, in Fig. 2 (A2), users can check the information of all categories under a rating by hovering on the bar and clicking to filter reviews by a rating. To meet the **DR4**, we designed a mouse-over triggered animation with the legend. By hovering on the legend, users can view the distribution across different ratings *horizontally* related to one category (Fig. 2 (A3)). The corresponding sectors of the selected category (e.g., ■ e1) in different pie charts will also be highlighted. In this way, users can easily check the distribution of one category across different user ratings. We added a smooth transition animation to make the changes between bar charts more natural. Moreover, users can also filter reviews by clicking the categories in the legend, such as filtering reviews that express extreme negative emotions (i.e., clicking ■ e1).

6 PROTOTYPE SYSTEM

To evaluate the proposed design with users, we built a prototype system with transparent design for booking hotels. We firstly collected the data of real hotels from one of the most popular platforms, Tripadvisor⁵. Then, we extracted the specific aspects of the three kinds of information from the

⁵<https://www.tripadvisor.com/>

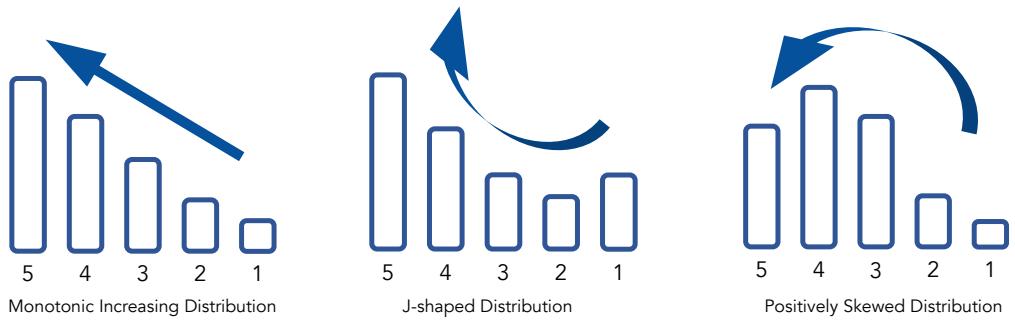


Fig. 3. Three kinds of distribution shapes for user ratings, observed from real data on Tripadvisor.

data automatically. Finally, we integrated the bias-aware design with data in a prototype system simulating the hotel-booking scenario online.

6.1 Data Collection

To evaluate our design effectively, we aimed to collect data from representative hotels online, which could mirror the opinions of ordinary users. Thus, we decided to select hotels from big cities in the world. Considering that participants of the formative study were all US-based, we retained this character in the user study. Hence, to reduce the impact of their previous impressions, we collected data from non-US cities that participants might not be familiar with. Two authors explored and compared the hotels' data in different cities. Considering that the COVID-19 may influence the user ratings and reviews of hotels online, we only collected user ratings/reviews in the time interval of 30 June, 2016 to 31 January, 2020 (before COVID-19 became International Concern ⁶). Indeed, the pandemic is likely to change demographics of travelers, their emotions, and the aspects they concern about a hotel. It would be interesting to explore and compare the changes of content and the potential biases in the content of user feedback before and after COVID-19. We leave this topic as part of our future work, because COVID-19 made travel extremely difficult, if not impossible, in most parts of the world when we conducted the experiments. Finally, we chose London as it had the most hotels as well as user reviews.

We initially crawled 976 hotels' pages under the tag "*London*" on Tripadvisor. To narrow down the scope and select representative hotels, we narrowed down the scope of hotels to 3-star hotels labelled by Tripadvisor as hotel-class because there are more 3-star hotels (40.7%) than others. These hotels had a price range from \$40 to \$200, with \$88 as the median price, and we found that hotels with price ranges from \$82 to \$105 contained more than 80% of the user ratings/reviews within the setting time interval. Therefore, we selected hotels based on this price range and filtered out the hotels with no feedback for the last six months. Finally, we got 57 hotels and crawled the data of these hotels, including the hotels' names, price, user ratings, reviews, etc. By controlling these variables, we could select the representative hotels and let users focus more on the user feedback to evaluate the bias-aware design effectively.

6.2 Data Distribution

After selecting the representative hotels by control variables, such as prices and the hotel-class, we explored the distribution shapes of user ratings in these hotels as the distribution can be affected

⁶The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January 2020, and a pandemic on 11 March 2020.

by the self-selection bias proven by previous works [49, 80]. We analyzed the user rating data of the 57 hotels and summarized three kinds of distribution in Fig. 3.

These three distributions include (1) the monotonic increasing distribution, (2) the J-shaped distribution, and (3) the positively skewed distribution, which are observed from the selected hotels' ratings on Tripadvisor. (1) The monotonically increasing distribution means that the shape of the user rating bars is monotonically increasing from 1 point to 5 points. (2) The J-shaped distribution is the typical distribution caused by self-selection bias as it has an extreme distribution with more 1-point and 5-point ratings [12, 49]. (3) The positively skewed distribution has more middle points (from 2 to 4 points) with fewer 5-point ratings compared to the other two distributions. This distribution is close to real-world data under lab conditions [61].

We randomly choose 15 hotels from 57 hotels with five for each distribution shape, considering that the distribution shape may affect people's perceptions of hotels, and their decisions and people usually narrow down to no more than 15 hotels for decision-making based on the pilot study. We further extracted and analyzed the user ratings and reviews for the selected 15 hotels, which is illustrated in the next subsection.

6.3 Data Processing

The data processing step consists of three parts corresponding to the three kinds of information mentioned in Section 4. We collected 5940 user ratings and reviews from the 15 hotels (MEAN = 368, MAX = 397, MIN = 320) in total between 30 June 2016 to 31 January 2020. We first obtained the reviewers' information (I1) based on the *reviewer badge* on Tripadvisor. Then we got the information about emotion (I2) and reported aspects (I3) using two automatic algorithms in the natural language processing (NLP) domain, including sentiment analysis and topic extraction. Based on these approaches, we mapped the information (I1-I3) into the bias-aware design.

6.3.1 I1: Reviewers' Rating Experience. We measured the reviewers' rating experience by separately calculating the number of total reviews they gave (I1-1) and the number of "helpful" votes they got (I1-2) on Tripadvisor. Then, we divided six categories for each kind of information referring to the reviewer badge on Tripadvisor. For example, we marked a reviewer as a "*Top Reviewer*" when they have more than 100 reviews and annotated a reviewer as a "*New Reviewer*" if he/she has only posted one review. For the second type (I1-2), referring to the classification of Tripadvisor, a reviewer with more than 100 "helpful" votes would be called a "*Top Contributor*" while a reviewer with 0 "helpful" votes is considered a "*New Contributor*". Hence, we can label each review by its reviewers' rating experience in these two aspects (each has six categories) and calculate the distribution of user reviews under these categories associated with user ratings.

6.3.2 I2: Sentiment Analysis of User Reviews. To extract the emotional polarity from textual reviews, we performed sentiment analysis on user reviews with AllenNLP [38]. It provides an off-the-shelf, ready-to-use sentiment analysis library based on Roberta [63] and the model was trained on the Stanford Sentiment Treebank [83] with a test accuracy of 95.11%. We divided the results into five categories ranging from -1.0 to 1.0 evenly and used one of the categories, including "*Positive Only*", "*Positive*", "*Neutral*", "*Negative*" and "*Negative Only*", to label the extremity of emotion in a user review. We obtained the distribution of the extremity of emotions associated with user ratings by counting the number of reviews in each category and their corresponding percentage.

6.3.3 I3: Reported Aspects of User Reviews. We used a topic extraction approach to extract keywords from user reviews as the reported aspects (I3). Firstly, we trained the topic extraction model using another hotel review dataset⁷ which consists of 878,561 reviews from 4333 hotels on

⁷<https://www.cs.cmu.edu/~jiweil/html/hotel-review.html>

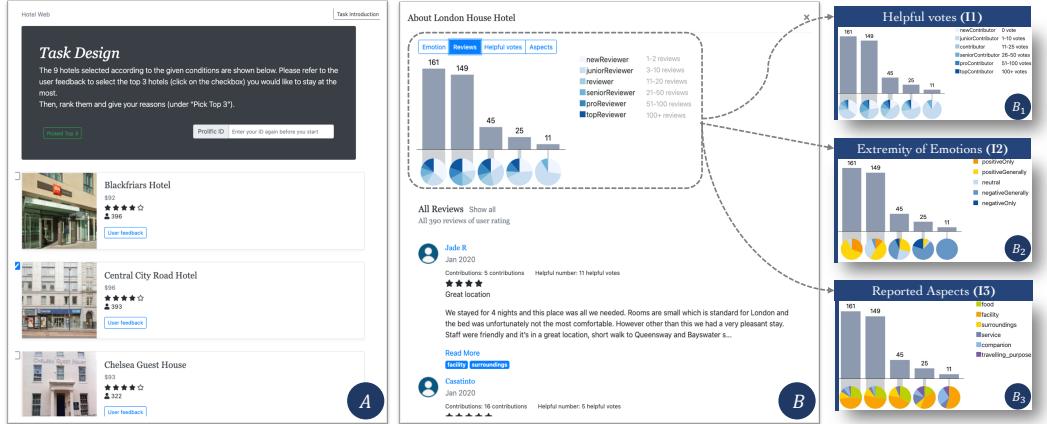


Fig. 4. Two main interfaces of the prototype website. (A) shows the list of hotels provided for participants to select from; (B) displays a detailed page of one hotel, including the bias-aware designs of user ratings. Users can switch the design with different types of information by clicking the navigation bar above.

Tripadvisor. Then, we applied the KeyBERT [43] model to extract keywords from reviews based on the pretrained model. We keep the 300 most frequent keywords from the extracted keywords and convert each of them to a vector by looking up its word embedding vector in GloVe. Through clustering these 300 vectors with K-Means algorithm, we found nine clusters.

Three authors analyzed the resulting clusters and excluded two clusters after discussion, whose keywords are the dates of user review (e.g., March or 15th). We also merged two clusters that are similar and are related to the environment of hotels. Finally, we got six categories that described different aspects of hotels in user reviews (i.e., food, facilities, service, surrounding environment, travel purpose, and companions). As one review may contain multiple categories, we calculated the percentage of a category by adding up the total number of reviews in all categories.

6.4 A Hotel-booking Website with Bias-aware Design

To evaluate the design with ordinary users, we simulated a hotel booking scenario by building a prototype website integrating the bias-aware design. To increase the scalability and user friendliness of the prototype, we implemented the basic functions of booking hotels and used a responsive framework that makes the interface adaptive to different screen sizes.

We first collected the interface design of many popular hotel booking websites on the Internet, such as Tripadvisor.com, Booking.com, Expedia.com, etc. We analyzed their interfaces and found that these websites use a similar framework of user interfaces, which shows the general information of each hotel in a card view and displays the detailed user reviews in a separate web page or a pop-up page. We adopted this design and integrated the bias-aware design into the user rating position in the detailed page of each hotel.

We show the two main interfaces of the prototype in Fig. 4, in which (A) shows the homepage with the introduction of the user study and a list of hotels. By clicking the button in a card view, users can open a new page in a pop-up window that displays the corresponding user ratings and reviews of one hotel in Fig. 4 (B). We display the bias-aware design with a navigation bar above, which is used to switch between different information types (Fig. 4 (B1)(B2)(B3)). There are four options in the navigation bar that include two types of rating experience of reviewers (i.e., written reviews (I1-1) and “helpful” votes (I1-2)), emotion (I2) and aspects (I3). Users can interact with the

design to filter user reviews below or check the distribution of one category across all different user ratings, as illustrated in Section 5.3.

We display the user reviews below the bias-aware design and show each review together with the user name, time, reviewer, user rating, and reported aspects, referring to the design of Tripadvisor. Users can scroll up and down to read user reviews in this page. To avoid gender bias, we replaced all the original reviewers' names with either abbreviations or random common names for both males and females.

7 USER STUDY

With institutional IRB approval, we conducted a between-subjects user study with 144 participants to evaluate the effectiveness of the bias-aware design. Our goal is to answer the general research question through the evaluation: **(How) does the bias-aware design help people make informed decisions?** We further decomposed the general research question into three subquestions.

- **RQ1:** Does the bias-aware design raise people's awareness of the self-selection bias compared to the baseline (the common design of user ratings)?
- **RQ2:** How do people use the bias-aware design to make decisions compared to the baseline?
- **RQ3:** How does the bias-aware design affect people's decisions compared to the baseline, if at all?

The goal of **RQ1** is to test the design in raising people's awareness of the self-selection bias compared to the baseline. In addition, to gain insights into improving our design and offering implications for future work, we observed and compared users' decision-making strategies in both conditions (**RQ2**). We set **RQ3** to investigate the effect of the bias-aware design on users' final decisions, and further verified that enhancing people's awareness of the self-selection bias could help people make more informed decisions.

7.1 Participants

We used the Prolific to conduct the user study online. To ensure that all participants had the experience of booking hotels online prior to our study, we set screening questions on the frequency and most recent activities of booking hotels. We initially recruited 144 participants and excluded the responses of eight participants who failed this quality control during the study.

We obtained responses from 136 participants after the quality check, of whom 68 (33 females, 34 males, and one prefer not to say) completed the experiment with our proposed prototype system, and the other 68 (35 females and 33 males) used the baseline system. The distribution of our participants' education background is as follows: undergraduate (46.3%), graduate (29.4%), technical/community college (8.1%), high school diploma (8.8%), doctorate degree (5.9%) and unknown (1.5%). The average age of the participants was 31 (Max= 73, Min = 20, SD = 9.97, 95% CI [29.14, 32.49]). Following the rules of the Prolific platform, we paid all participants at a rate of £7.60/hour. The actual amount for each individual depended on the task completion time and the Prolific settings according to our predefined time range (30 min).

7.2 Experiment Setup and Task Flow

7.2.1 Baseline System. For comparison, we created a baseline system by removing our visual design but keeping other hotel booking-related functions available in the prototype system. The baseline system shared the same home page (Fig. 4 (A)) with our proposed system, but it adopted a simpler visualization of user ratings in the detail page, as shown in Fig. 5 (C), on which users could filter reviews by semantic tags or rating bars, like the common hotel booking sites used in daily

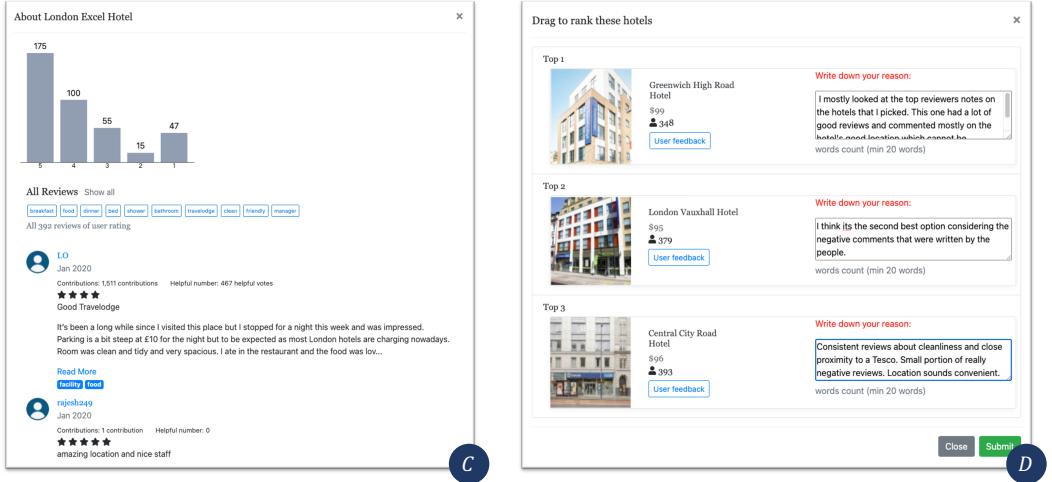


Fig. 5. The interface provided by the baseline system is shown in (C). The interface in (D) is used to record the results of user selections and the reasons for uploading in both prototype and baseline systems.

lives. We hosted both systems on a web server that enables participants to gain access via a public link.

7.2.2 Configuration of the Website. To reduce the confounding effect of irrelevant variables in the study, we controlled the information presented to make participants perceive a hotel primarily by viewing user ratings and reviews. We preprocessed the information presented to participants on both websites, such as the price, number of user reviews, and user ratings of each hotel, within a narrow range, as described in Section 6.1. During the study, the candidate hotels were presented in a randomized order for each participant to avoid anchoring bias. Moreover, we kept the page's functionality simple. We only used photos of the hotels' entrances or buildings to reduce any influence of photos on the participants' decisions.

7.2.3 Task Flow. We designed a task for user study that simulates the real-world online hotel booking experience. To complete the task, the participant had to go through the following three steps.

- (1) They first watched an introductory video⁸ (1-2 minutes) about the usage and task of the experimental websites. Through the video, we let the participants imagine that they were required to choose a hotel for a vacation in London based on user feedback, while paying less attention to other information controlled in our study, like price or hotel class.
- (2) Then, participants browsed a pool of 15 candidate hotels (identical for the experimental and baseline conditions), including the user ratings and reviews in the given system and shortlisted the top three hotels they may want to make a reservation at. To compile the shortlist, the participants needed to specify three hotels by checking the box next to the hotel name on the homepage (Fig. 4 (A)). The choice of hotel depended totally on the participant's personal judgment.
- (3) After that, participants could rank the selected hotels by dragging the hotel up or down in a pop-up window (Fig. 5 (D)). They needed to justify their decisions in the text area of each

⁸We upload the introduction video for the experiment in the supplementary materials

hotel. Note that their draft answers in the pop-up window were automatically saved, so they could return to the previous page to view the details of a hotel at any time if needed. To conclude the task, the participants had to submit their final choices and feedback by clicking the *Submit* button in Fig. 5 (D).

These three steps allowed us to examine the effectiveness of our design and collect information about how people make decisions within the two systems.

7.3 Pilot Study

Before launching the formal experiment, we conducted a pilot study to test our task design and configuration with eight participants (four females), who were undergraduate or postgraduate students recruited from a local university. Their ages ranged from 21 to 27 (Mean = 24, SD = 2.51, 95% CI [21.90, 26.10]), and their backgrounds included computer science, finance, psychology, and biology. These participants were randomly divided into two groups; half completed the task with our prototype, while the rest used the baseline system. After finishing the task, they filled out a post-study questionnaire and participated in a brief interview to share their study experiences.

Based on the feedback from the pilot study, we adjusted several designs in both conditions. Seven out of eight participants (87.5%) reported that they usually inspect and compare fewer than 10 hotels in detail in real life, because it became overwhelming and time-consuming beyond that number. Examining 15 hotels fully deviated from their common practice and they were not able to remember and compare the information of all these hotels at once. Thus, we reduced the number of hotels in the candidate pool from 15 to 9 for our formal study by randomly removing two hotels from each distribution type mentioned in Fig. 3. Furthermore, as informed by the pilot study, we further streamlined our study procedure (presented in the next subsection) and enriched the details of the instruction videos.

7.4 Procedure

In our formal study, 144 participants were randomly split between the baseline group and the experimental group. All participants first signed a consent form, agreeing to join the experiment and allowing us to collect basic information about them. We recorded their operations during our study and obtained their demographic data from the Prolific platform. Then, following the three steps presented in the task flow (Section 7.2.3), each participant watched an introduction video, browsed the candidate hotels, specified the top three choices, wrote down the reasons for the choice, and submitted in the prototype system. After sending their selections and reasons, each participant filled out a post-study questionnaire, which assessed their awareness of potential bias and collected qualitative feedback on our design and systems (RQ1). The average time for the whole study was 34 minutes (baseline: 29 minutes, prototype: 39 minutes). We did not control the time of our user study because we aimed to simulate the real-world scenario of booking hotels so that the study results could truly reflect people's decisions and their decision-making process. We concerned that strictly controlling the time may add stress to the participants and affect their behavior during the experiment and consequently the observed results.

7.5 Operational Data Collection and Analysis

General quality check. We carried out a general quality control by checking the usage time of each participant. The time started counting from when they opened our prototype and ended by submitting their feedback to the questionnaire. They can get the completion code and return it to Prolific after successfully finishing all questions in the questionnaire. Eight responses (out of 144) were removed due to the extremely short time recorded (i.e., less than 1 minute for each hotel on

average, set based on the pilot study) or they did not return the completion code of Prolific. Finally, we got 136 valid responses to the questionnaire (68 responses for each condition).

Quality check of interactive operations. To analyze and compare users' behaviors in both conditions, we implemented several event listening functions in our prototype to collect participants' operations during the study (RQ2). We collected all records of participants' interactive behavior for browsing and selecting from nine candidate hotels. The behavior logs include the basic interactions of all participants, including clicking, hovering, and scrolling up or down in our system. We counted and averaged the number of user interactions, i.e., clicks and hovers, on each individual rating bar. The interactions recorded with the user rating bars in the prototype include those on the pie chart linked to each bar (Fig. 2).

To ensure the completeness of the data logs of participants' operations (e.g., click, scroll, hover, etc.) on the study interface (RQ2), we defined a threshold of the minimum number of interactions required for each participant to check the information of all 9 hotels provided in the study. This threshold, 102 times of operation, is set based on the observation of pilot study. After scrutinizing the operation data using the threshold, we identified 26 behavior logs (out of 136) that failed the quality check. We found that these participants all had high credit scores on Prolific and their operational logs are incomplete because of the lack of hovering operation records, which indicated that they might not fail the check on purpose. To figure out the reasons, we tried to get in touch with these 26 participants via the Prolific chat. Through multiple rounds of communication, we confirmed that these participants failed the quality check due to the compatibility issues occurring with older browsers. Therefore, we only removed their operation logs from the data analysis while keeping their feedback on the questionnaire. Finally, we retained 136 valid responses to the questionnaire and 110 behavior logs (53 with the baseline, 57 with our design) from participants.

8 RESULTS

In this section, we structure the results according to the research questions mentioned in Section 7 (RQ1-RQ3). We first provide the results of the questionnaire reflecting participants' awareness of self-selection bias when making decisions (RQ1). We next show and compare the strategies used by the participants to make decisions in the two conditions of the study (RQ2). We also summarize patterns in participants' operational logs in both conditions to reflect how the experimental group used our prototypes and compare their actions to those in the baseline condition (RQ2). Finally, we present and compare the final selections between the two conditions (RQ3).

8.1 Raising Awareness of the Self-selection Bias (RQ1)

We measured people's awareness of the self-selection bias in user ratings and reviews by the post-study questionnaire. Note that the self-selection bias is implicit, which needs to be measured carefully without revealing the goal of our study (testing awareness) during the experiment. To effectively test whether the bias-aware design works or not, we instructed each participant to book a hotel using the given prototype system under a concrete scenario (going to London on vacation) as how they would do it in the real life. We followed a psychological method called implicit association test (IAT) [73] and adapted the questionnaire of a famous IAT project called Project Implicit⁹ to design our questions for accessing awareness. IAT is widely used to assess people's implicit bias, stemming from attitudes or stereotypes that mostly occur outside of people's consciousness and control [73].

As shown in Fig. 6, we present our five adapted IAT questions (Q1–Q5) with the corresponding participant ratings that measure their overall awareness (Q1) and the awareness of the three types of

⁹<https://implicit.harvard.edu/implicit/education.html>

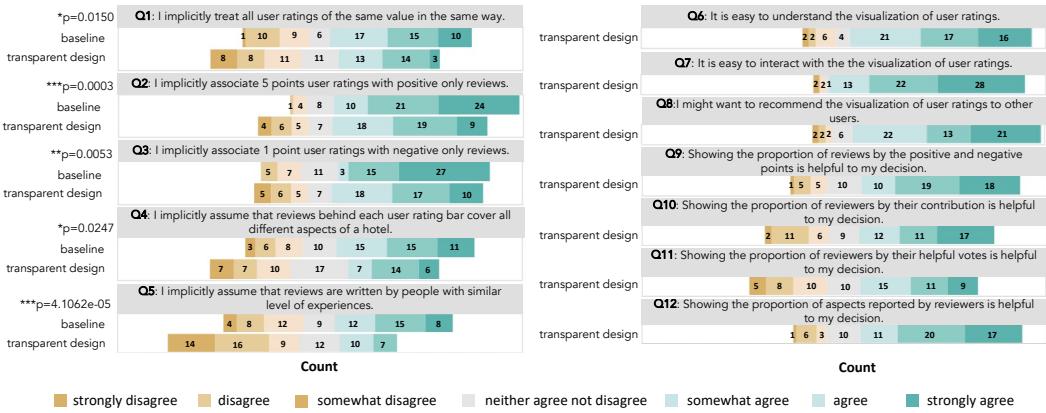


Fig. 6. Participants' ratings of the awareness testing questions in post-questionnaire and the subjective response to our design in the experimental group.

information (Q2-Q3: emotion [I2], Q4: aspects [I3], Q5: reviewers [I1]) caused by the self-selection bias. We used the Mann-Whitney test [66] to compare the result of both conditions and calculated the 95% bootstrap confidence intervals for each question [32]. We display the responses of other post-study questions (Q6-Q12) that evaluate the information provided by the transparent design as well as its usability on the right side of Fig. 6.

The results of Q1–Q5 show that people in the experimental condition disagreed more with the descriptions, which indicated that they were significantly more aware of the self-selection bias in user ratings and reviews. Q1 measures the general awareness of participants between two conditions, which indicates that people booked hotels with our design were more aware of the diverse information behind user ratings. Among these cases, participants were more aware of the diverse experience of reviewers behind user ratings (Q5, p -value < 0.001) more than other cases using the transparent design. This result echoes the last strategy used by participants when selecting hotels in Section 8. For the responses of Q4, participants' awareness of the different aspects of a hotel behind user ratings (p -value = 0.0247) had the least difference between the two conditions. The reason is that participants using the baseline could also view the tags of diverse aspects for filtering reviews under the user rating as the ordinary hotel booking websites. Q2 and Q3 evaluate the level of participants' awareness of the extreme emotions behind user ratings, including the positive and negative emotions. Participants who used the transparent design were relatively more aware of the negative emotions (Q3, p -value < 0.01) than positive ones (Q2, p -value < 0.01).

We collected participants' feedback on the information and usability of our design (on the right side of Fig. 6) and found that our design was perceived to be easy to understand (Q6, Mean = 5.28, SD = 1.51, 95% CI [4.92, 5.64]) and interact with (Q7, Mean = 5.96, SD = 1.30, 95% CI [5.64, 4.26]). In general, participants would like to recommend our design to other people (Q8, Mean = 5.46, SD = 1.48, 95% CI [5.12, 5.82]). We also asked the participants who used the transparent design about the helpfulness of the information provided (Q9–Q12). Overall, participants agreed that showing the proportion of aspects (Q12, Mean = 5.24, SD = 1.63, 95% CI [4.84, 5.61]) and emotions (Q9, Mean = 5.24, SD = 1.62, 95% CI [4.86, 5.64]) behind user ratings is useful for their decision-making. Showing the information about the reviewers, such as the number of reviews they have published (Q10, Mean = 4.75, SD = 1.89, 95% CI [4.29, 5.18]) and the number of “helpful” votes those reviews received (Q11, Mean = 4.34, SD = 1.81, 95% CI [3.90, 4.75]) are deemed relatively less useful than

Table 1. The coding scheme and summary of strategies from the participants in both conditions.

Strategies	Condition	Percentages
Checking the Specific Aspects in User Reviews	Prototype	88.45%
	Baseline	94.23%
Considering Positive and Negative Reviews Collectively	Prototype	55.77%
	Baseline	28.85%
Referring to the User Rating Distribution	Prototype	13.46%
	Baseline	48.08%
Referring to the Number of Negative Reviews	Prototype	17.31%
	Baseline	13.46%
Referring to the Reviewers behind User Ratings	Prototype	23.08%

other information. Several participants reflected that they got used to filtering reviews by positive or negative aspects rather than using the information related to the reviewers.

Above all, we observed that participants are more aware of the potential bias after viewing data distribution behind user ratings with the help of the bias-aware design than those in the baseline condition. They were more aware of the difference of reviewers' experiences than other kinds of information based on the responses of questionnaire. In addition, participants who experienced our design gave positive feedback on its usability and usefulness. Most participants in the experimental group reported that the distribution of extreme emotions behind user ratings was the most useful information for them to make decisions.

8.2 Strategies for Choosing Hotels (RQ2)

To understand how people leverage our design to make decisions, we asked participants to share their strategies for selecting, eliminating and ranking hotels in the post-questionnaire. Together with the uploaded reasons in the systems (Fig. 5 (D)), two authors coded the strategies using thematic analysis [17] and finally reached consensus through two rounds of discussion. In general, we found that participants in both conditions share similar strategies of choosing hotels while they put different emphasis on several aspects. We extracted five typical strategies from their feedback and summarized them in Table 1. We calculated the percentage of every strategy among all participants in each condition; four out of the five strategies occurred in both. We derive and report four salient findings regarding these strategies below.

8.2.1 Most participants in both groups selected hotels by checking the detailed aspects in user reviews. As shown in Table 1, 88.45% and 94.23% of participants in the two groups respectively mentioned they would like to inspect certain aspects of personal interests in the detailed user reviews when booking hotels. For example, P20 selected a hotel by checking several aspects she cared about.

“Cleanliness seems to be a consistent theme. I filtered reviews on cleanliness and looked for ones with the least amount of really negative reviews on this matter.” – P20 (F, 40, Prototype)

P5 chose a hotel by checking the options for breakfast.

“Breakfast included vegetarian options. Breakfast is my favorite meal and I would not want to stay in a place where I would have to buy breakfast outside of the hotel.” – P5 (M, 25, Baseline)

The difference is that participants who used our prototype system could filter the detailed aspects of reviews corresponded to the user rating bars while the participants in the baseline condition used tags to filter user reviews.

8.2.2 People tend to consider both positive and negative reviews together when using the prototype system. The summary of strategies shows that participants who used our design were more likely to consider both positive and negative reviews (55.77%) compared to those in the baseline group (28.85%). This provides evidence that our design can help users gain a comprehensive view of others' feedback on a hotel. For example, P5 chose Hotel 6 (J-shaped) and gave the following rationale:

"I choose this hotel because the good to bad reviews ratio is optimal. The few bad reviews don't seem to bad to me." – P5 (F, 20, Prototype)

P28 only used negative feedback to judge a hotel.

"Since the prices were very similar, I strictly chose hotels based off the most negative reviews and the distribution of ratings." – P28 (M, 33, Baseline)

However, in the fourth row of the Table 1, participants using the prototype referred to the number of negative reviews a little bit more than those using the baseline ($17.31\% > 13.46\%$). The reason is that participants were interested in the transparent information of the 1-point rating as they interacted more with the 1-point bar as shown in the Fig. 7 (B).

8.2.3 People rely less on the rating distribution with the transparent design than those with the baseline. Although participants' selection of hotels in each group were affected by the rating distribution according to Fig. 7, in their feedback of strategies, we found that they have different degrees of dependence on the rating distributions. We could see that 48.08% of participants filtered hotels by taking the user rating distribution into account in the baseline system, while the participants less mentioned the user rating distribution using our system (13.46%) (Table 1). For people who chose hotels with the transparent design, some of them mentioned the rating distribution when they were aware of the potential bias. For instance, P62 (in the prototype group) mentioned that he eliminate the hotels with more 5-point ratings and 1-point ratings.

"I eliminate hotels with U shape of comments or the ones have unacceptable negative comments." – P62 (M, 28, Prototype)

However, participants used the baseline relied on the rating distribution by considering either positive or negative reviews, such as P65 in the baseline group.

"If there's a high proportion of 5- and 4-star ratings, I am much more reassured that my experience will be good." – P65 (F, 36, Baseline)

The reason is that people can view the overall content of user reviews through the transparent information presented in our design. Thus, users are less likely to judge the quality of a hotel **solely** based on the rating distribution.

8.2.4 The participants prefer to rely on the feedback of professional reviewers when choosing a hotel using the transparent design. We discovered that 23.08% of participants in the experimental condition (Table 1) used a new strategy not available in the baseline system; that is, taking the feedback from professional reviewers into consideration. P41 (in the prototype group) mentioned that

"I chose these three hotels generally because they had the fewest one-star ratings given by the 'top' and 'pro' reviewers. I trust these peoples ratings more as they have

experience giving reviews and consistently do so and are not likely to just give a bad rating based on one anomalous trip.” – P41 (M, 33, Prototype)

8.2.5 Summary of Operational Logs. To explore how participants used our prototypes, we analyzed and compared the operational logs of all participants in both conditions. We plotted the result in Fig. 7 (B) and (C). As shown in Fig. 7 (B), the average number of clicks indicates how frequently users filtered reviews, while the average number of hovers implies how often they viewed relevant information by putting the cursor over it. Participants were more inclined to filter negative reviews in both systems and, compared to the baseline, participants who used our design hovered more times over the design bars as well as pies to view the information. We explored the possible reasons for this phenomenon from the responses in the post-study questionnaire, and found that participants who used our design preferred to get a general impression of a hotel by viewing the transparent information instead of reading a lot of user reviews.

“I really want to use this design in the future as it will really help, firstly, not to waste your money, and secondly not to waste your time in searching for good hotels.” – P11 (F, 28, Prototype)

The above findings are also confirmed by the average number of different kinds of interactions across the whole web site during the experiment (Fig. 5 (C)). More specifically, the total number of clicking and hovering interactions with our system far exceeded those of the baseline, while the number of scrolling up or down actions was considerably lower. This indicates that the participants with our design read fewer reviews than those in the baseline system, as scrolling the web page is necessary to see more user reviews that cannot be displayed within one screen. Note that the system interfaces are adaptive, so there is no need for users to scroll the page when looking at user ratings. Participants in the experimental condition also acknowledge in their post-questionnaire responses that they felt that our bias-aware design was helpful for them to locate more relevant reviews conveniently.

To sum up, we found that participants in the experimental group used relatively different strategies for referring user ratings compared with the baseline condition. They tended to consider the information from different ratings more comprehensively and paid relatively less attention to any single type of feedback (e.g., negative feedback). Furthermore, participants in the experimental condition seemed to trust the feedback from professional reviewers more than other reviewers when they were informed about the distribution of reviewers’ experiences behind user ratings.

8.3 Selection Result (RQ3)

We summarize the participants’ final selections of hotels in Fig. 7 (A). The bar chart shows the percentage of each hotel selected in both conditions respectively. We grouped the hotels by the distribution type of user ratings (Fig. 3) and compared the participants’ selections in the two conditions. We distilled several findings from participants’ selection reasons by coding and analyzing the reasons provided by the participants using thematic analysis [17].

8.3.1 Participants’ choices of hotels were affected by the distribution of user ratings. As shown in Fig. 7 (A), the hotels with a monotonic increasing distribution of user ratings got the most votes, while hotels with a positive skewed distribution of user ratings were the least favorable to our participants. The hotels with a J-shaped user rating distribution gained moderate votes. Therefore, people’s decisions could be affected by the distribution of user ratings as the participants tended to choose hotels with more 5-point and fewer 1-point ratings.

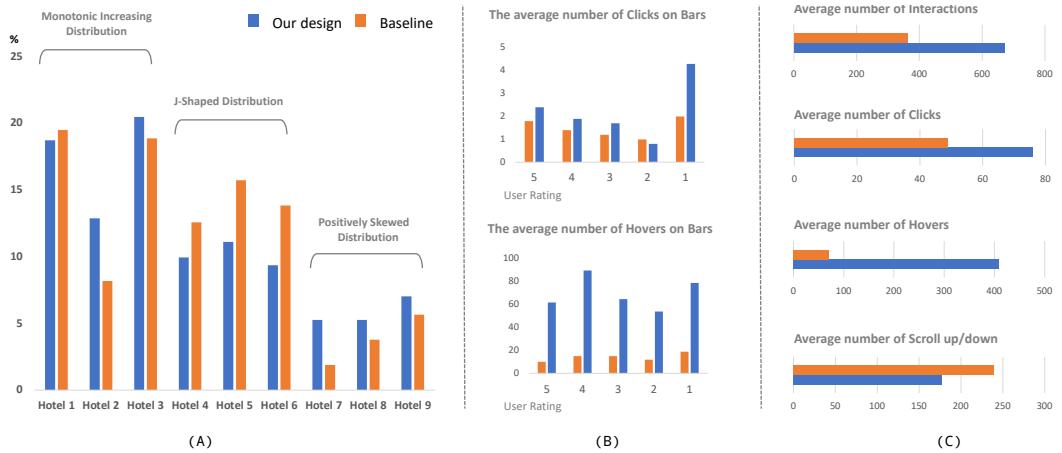


Fig. 7. The statistical record of participants' choices of hotels and their interactions in the study. (A) shows the total number of times each hotel was selected by participants; (B) shows the average number of interactions on each rating bar for each participant in two systems. The x-axis represents the user ratings from 5 points to 1 point. The x-axis of (C) displays the average number of different types of interaction for each participant in two conditions.

8.3.2 Seeing the transparent information behind the user ratings could affect the selection of hotels. The other finding is that, in Fig. 7 (A), less percentage of participants using the bias-aware design chose the No.4 - No.6 hotels under the J-shaped user rating distribution (the typical biased distribution caused by the self-selection bias) than those using the baseline. Moreover, more participants chose the hotels with a positively skewed rating distribution (the distribution with the least bias [61]) using the prototype system than those with the baseline. We analyzed the reported reasons for selection elicited by the participants. We discovered that, compared to people applying the baseline system, participants in the experimental condition considered the general rating distribution more when they could see the transparent and more detailed information behind different ratings (1–5 points). To be more specific, they would like to inspect and analyze the auxiliary data shown in the pie charts under the user rating bars (Fig. 4 (B)) and make decisions based on it.

“I selected hotels according [to] the most balanced of all aspects of ratings to reviewers and contributors in the pies.” – P12 (M, 30)

Generally, compared with the baseline, a higher percentage of participants in the experimental group selected hotels with a less biased distribution of user ratings (i.e., positively skewed distribution [61]), and a smaller percentage of them chose the hotels with the potentially biased distribution (e.g. J-shaped distribution).

9 DISCUSSION

The user study results indicate the effectiveness of our design for raising awareness of the self-selection bias and facilitating people's decision-making process. In this section, we discuss lessons learned from the implementation of design and user study. We also reflect on the implications for supporting people's decision-making by raising awareness of the self-selection bias through an informed and unintrusive way.

9.1 Effect of Information Transparency

We decided the exact transparent information by a formative study that collected the critical information people care about and people's perceptions of the transparent information. In the study, we showed the information combining user ratings with reviews to people and explore how they exploit the information and make decisions. The results indicate that the information provided are helpful for decision-making in general (Fig. 6) while changing people's behaviors and their strategies of referring user ratings and reviews for decision-making (Section 8).

However, it is necessary to mention that people in the group with the bias-aware design spend 10 more minutes on average than the baseline in the study. The reason was that, on the one hand, the experimental group needed to watch 1 more minute of the introduction video to understand our design and task before the study (Section 7.4), and they answered more questions in the post-questionnaire. On the other hand, the operation logs shows that they interacted more with the bias-aware design than the baseline (Fig. 7). Hence, how much and the granularity of transparent information to be shown are all needed to be considered. If we disclose too much information, users may abandon checking the details as they can be overwhelmed. For example, while other aspects of user ratings and reviews may be biased (e.g., reviewers' gender and ethnic group), the formative study results suggest that they are not the key concerns of most crowd respondents during hotel selection. Rather than presenting all potentially biased dimensions of the user-generated data, we only emphasize on the most salient ones that have strong impact on people's decisions.

In addition, the process of making information transparent is not just about the information itself. It also concerns how to present the information that was previously "invisible" to users without overloading users. It is thus necessary to consider and reflect the information needs of users in the design. We investigated how to present transparent information intuitively based on the formative study and the discussion with two visualization experts. Then, we proposed three alternatives and compared how people perceive and explain the information in these designs. While some of the alternatives could better support visual analytic tasks, they are more visually complex, more challenging to interact with, more space-consuming, and harder to scale. Based on user feedback in the pilot study, we selected the one design that is the most acceptable to our target users (i.e., lay public) and improved it by adding user-friendly interactions. This process ensured that at least most ordinary users could easily learn to use our design and grasp the conveyed information quickly and freely in their decision-making process. In general, designs related to information transparency should always consider the needs and characteristics of users and strike a balance between information transparency and information complexity.

9.2 Evaluation of Awareness

Certain biases in user-contributed contents (e.g., imbalance in gender) could be alleviated by applying a better sampling technique proactively, such as the methods used by previous works [1, 6, 61, 71, 95], as discussed in the Related Work (Section 2.2). However, some other types of biases in user-generated data are caused by unconsciousness, such as people's extreme emotions [81], their knowledge/expertise [44], and selection of individuals [12]. They tend to be more implicit and may not be easily detected, measured, and mitigated by algorithms. Therefore, in this work, we chose one kind of implicit bias (i.e., self-selection bias) and proposed to raise people's awareness of the bias as the goal of our work [7], on the assumption that such biases could not be technically eliminated from the data.

One challenge we met in this work is the evaluation of the people's awareness, as it is not proper to directly ask whether they are aware of the bias or not. A previous study by psychologists has shown that people tend to appear knowledgeable when their awareness is explicitly tested [84]. We

explored the literature and found that the way of evaluating the awareness should be associated with the concrete task and goal. For example, a recent work by CSCW researchers used the reliability scores in the pre- and post-survey to measure the awareness of the echo chamber effect [51]. Another work by Eslami et al. [35] tested users' awareness of the bias in algorithms by labeling whether they could articulate a discrepancy between their intended review score and the system output. We designed the questions in the post-questionnaire based on the implicit association test (IAT) in psychology [73] and the questions are adapted with the defined transparent information related to the self-selection bias. Therefore, the method of evaluation of people's awareness should be carefully considered and designed, especially for the awareness of the implicit bias.

Additionally, there is an interesting finding of the awareness testing result in Fig. 6. People were more aware of the potential biased reviewers' experience than the extreme emotions and various aspects in user reviews while the emotion is the most care than the other two types of information according to the formative study (Section 4). We attribute this to people's belief updating problems as people have a demand of the extreme user reviews (e.g., negative reviews) to some extent for making decisions [2]. Hence, although people may be aware of the bias in the extreme feedback, they still tend to associate the user rating score (e.g., 1-point) with the corresponding emotions (negative only). Moreover, we also attribute this to one limitation of our work in the evaluation of people's awareness. The questions for evaluating awareness in the post-questionnaire could be improved by gathering people's reactions and perceptions of the aware-testing questions in the pre-study session or pilot study. Therefore, evaluating people's awareness of bias needs to consider implicit question design, specific application scenarios and pre-estimate people's responses to the evaluation method.

9.3 Unintrusive Design for Awareness-raising and Informed Decisions

In our design process, we decided to use a unintrusive way to raise people's awareness of the self-selection bias. The reason is that the direct warning of potential bias that explicitly informs people can be a kind of inducement for users which may cause unwanted emphasis on the data rather than the decision [57]. Therefore, we did not add any explicit prompts or warnings of the self-selection bias in the design, and we even did not reveal the goal of our study or design during the experiment. The study results show that this implicit approach can effectively increase people's knowledge and perception of possible biases in user ratings and reviews without distracting them from their decision process during the experiment. However, in the study results (Fig. 6), at least for the awareness of the different aspects (I3), it seems that people's general level of awareness has not reached the ideal level. One reason may be the implicit way of the evaluation and the design. Hence, it deserves further exploration of the effects of implicit design, as well as how to increase awareness in the implicit way.

In addition, it remains to be studied on how to efficiently offering the awareness-raising design for informed decisions in daily life. The latest research in the visualization domain proposed a method to help people realize their biased behavior during the data analysis process by showing the interaction logs explicitly in visual analytic systems [72, 91]. We could adopt this way in the online decision-making scenario by making people's behaviors transparent (e.g., scanning extreme reviews) and informing them about the potentially biased behavior on the fly. From the perspective of applying visualization, the literacy of visualization is key point that should be considered by designers as ordinary users may need some efforts to understand and interpret the information of the design.

9.4 Limitations and Future Work

We acknowledge that this work has several limitations. One limitation is that the sentiment analysis results and the keywords from user reviews involved in our experiment were all processed by automatic algorithms. Even though we chose a widely used tool [38] and a powerful model [43], we still cannot guarantee the algorithms are unbiased. These are emergent tools for checking model fairness [10], and we hope to ensure the use of unbiased algorithms with these tools before integrating them into the backend of our system in the future.

Secondly, for the purpose of our controlled study, variables other than reviews and derived information are confined to a narrow range (e.g., price) to limit potential confounding effects. In a real world situation, users may face a wide variety of available hotels and their choice may be largely influenced by other factors, such as price, location, and even images of the hotel rooms. Moreover, considering the duration of the experiment and the burden on users, we only provided nine alternative hotels, while people may browse much more candidates in reality. We consider assessing people's awareness of the self-selection bias in more complex situations, while using more hotels to evaluate the design in future work.

Thirdly, as we conducted our study using an online crowdsourcing platform, there are some uncertainties in this setup, such as unstable internet connections, browser incompatibilities, and other technical problems. To reduce the impact of this, we implemented some detailed functions to optimize the user experience, such as the adaptive interface design, prompt windows in the system, etc. While these helped in the experiment, several participants said they experienced compatibility and network issues. Future work needs to provide a more robust system for evaluating the design with a broader audience. Additionally, a long-term study is also necessary for observing how users leverage the bias-aware design for making decisions based on online ratings/reviews in real situations.

Despite these limitations, the user study does help us to learn more about users and their experiences, identify possible shortcomings in our design, and pinpoint improvements for our future work. In addition, exploring the impact of COVID-19 on the user-generated content and the possible bias behind (e.g., ratings and reviews of products or services) could be an interesting future direction to explore. We will also explore if our proposed approach can be generalized to other real-life scenarios in which user-generated content may affect users' decisions (e.g., buying a product online) in the future. As such, we will continuously improve our work that aims to raise people's awareness of various biases, and reflexively apply the design in other scenarios (e.g., buying products) based on the broader HCI and CSCW community's feedback.

10 CONCLUSION

We proposed a bias-aware design for user ratings with the aim of raising people's awareness of the self-selection bias. The design shows the proportions of three kinds of information that are related to the self-selection bias while affecting people's decision-making when referring user ratings/reviews. To evaluate whether the design could increase people's awareness and help them make decisions, we conducted an online study through a crowd-sourcing platform with 136 participants. The results show that the bias-aware design can significantly increase people's awareness compared to the baseline, and people can use the design more efficiently through interaction, thus facilitating their decision-making process. We analyzed several key points obtained from this work, including transparency information, evaluation of people's awareness of bias, and the non-intrusive design. Future work may further explore these points under various tasks or scenarios of decision-making. We hope this work can inform and inspire designers and researchers in the broad HCI communities to investigate bias from the end user's perspective.

11 ACKNOWLEDGEMENT

This work is partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under General Research Fund (GRF) with Grant No. 16203421.

REFERENCES

- [1] Kemal Kivanç Aköz, Cemal Eren Arbatlı, and Levent Celik. 2020. Manipulation Through Biased Product Reviews*. *The Journal of Industrial Economics* 68, 4 (2020), 591–639. <https://doi.org/10.1111/joie.12240> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/joie.12240>
- [2] Sandro Ambuehl and Shengwu Li. 2018. Belief updating and the demand for information. *Games and Economic Behavior* 109 (2018), 21–39.
- [3] Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive Opinion Summarization in Quantized Transformer Spaces. *Transactions of the Association for Computational Linguistics* 9 (2021), 277–293.
- [4] Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. 2015. Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 1803–1808.
- [5] Sinan Aral. 2014. The problem with online ratings. *MIT Sloan Management Review* 55, 2 (2014), 47.
- [6] Georgios Askalidis, Su Jung Kim, and Edward C. Malthouse. 2017. Understanding and overcoming biases in online review systems. *Decision Support Systems* 97 (2017), 23–30. <https://doi.org/10.1016/j.dss.2017.03.002>
- [7] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.
- [8] Elias Bareinboim and Judea Pearl. 2012. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*. PMLR, 100–108.
- [9] Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6331–6336.
- [10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.
- [11] Elisa Bertino, Shawn Merrill, Alina Nesen, and Christine Utz. 2019. Redefining Data Transparency: A Multidimensional Approach. *Computer* 52, 1 (2019), 16–26. <https://doi.org/10.1109/MC.2018.2890190>
- [12] Bharat Bhole and Brid Hanna. 2017. The effectiveness of online reviews in the presence of self-selection bias. *Simulation Modelling Practice and Theory* 77 (2017), 108–123. <https://doi.org/10.1016/j.simpat.2017.05.005>
- [13] Markus Binder, Bernd Heinrich, Mathias Klier, A. Obermeier, and Alexander Schiller. 2019. Explaining the Stars: Aspect-based Sentiment Analysis of Online Customer Reviews. In *ECIS*.
- [14] Todd Bishop. 2015. Amazon changes its key formula for calculating product ratings and displaying reviews. *GeekWire* (June 20), <http://www.geekwire.com/2015/amazon-changes-its-influen-tial-formula-for-calculating-product-ratings> (2015).
- [15] Eivind Bjørkelund, Thomas H. Burnett, and Kjetil Nørvåg. 2012. A Study of Opinion Mining and Visualization of Hotel Reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services* (Bali, Indonesia) (IIWAS '12). Association for Computing Machinery, New York, NY, USA, 229–238. <https://doi.org/10.1145/2428736.2428773>
- [16] Erik Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. 2007. Automatic Sentiment Analysis in On-line Text.. In *ELPUB*. 349–360.
- [17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.
- [18] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf>
- [19] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3995–4004.
- [20] Yung-Chun Chang, Chih-Hao Ku, and Chun-Hung Chen. 2019. Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management* 48 (2019), 263–279. <https://doi.org/10.1016/j.ijinfomgt.2017.11.001>

- [21] Swagato Chatterjee. 2020. Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach. *International Journal of Hospitality Management* 85 (2020), 102356. <https://doi.org/10.1016/j.ijhm.2019.102356>
- [22] Y. Chen, L. Chen, and Y. Takama. 2015. Proposal of LDA-Based Sentiment Visualization of Hotel Reviews. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. 687–693. <https://doi.org/10.1109/ICDMW.2015.72>
- [23] Judith A. Chevalier, Yaniv Dover, and Dina Mayzlin. 2018. Channels of Impact: User Reviews When Quality Is Dynamic and Managers Respond. *Marketing Science* 37, 5 (2018), 688–709. <https://doi.org/10.1287/mksc.2018.1090>
- [24] Simona Cicognani, Paolo Figini, and Marco Magnani. 2016. Social influence bias in online ratings: a field experiment. (2016).
- [25] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic generation of natural language explanations. In *Proceedings of the 23rd international conference on intelligent user interfaces companion*. 1–2.
- [26] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18, 5 (2008), 455.
- [27] Khalil Damak, Sami Khenissi, and Olfa Nasraoui. 2021. Debiased Explainable Pairwise Ranking from Implicit Feedback. In *Fifteenth ACM Conference on Recommender Systems*. 321–331.
- [28] Bart De Langhe, Philip M Fernbach, and Donald R Lichtenstein. 2016. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research* 42, 6 (2016), 817–833.
- [29] Peter De Maeyer. 2012. Impact of online consumer reviews on sales and price strategies: A review and directions for future research. *Journal of Product & Brand Management* 21 (04 2012), 132–139. <https://doi.org/10.1108/10610421211215599>
- [30] Gianluca Demartini, Kevin Roitero, and Stefano Mizzaro. 2021. Managing Bias in Human-Annotated Data: Moving Beyond Bias Removal. *CoRR* abs/2110.13504 (2021). arXiv:2110.13504 <https://arxiv.org/abs/2110.13504>
- [31] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.
- [32] Thomas J DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical science* 11, 3 (1996), 189–228.
- [33] Jiahua Du. 2020. *Advanced Review Helpfulness Modeling*. Ph.D. Dissertation, Victoria University.
- [34] Nico Ebert, Kurt A. Ackermann, and Bjorn Schepler. 2021. Bolder is Better: Raising User Awareness through Salient and Concise Privacy Notices. (2021).
- [35] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. “Be careful; things can be worse than they appear”: Understanding Biased Algorithms and Users’ Behavior around Them in Rating Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
- [36] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI ’19). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300724>
- [37] Jonathan St BT Evans and Keith Ed Frankish. 2009. *In two minds: Dual processes and beyond*. Oxford University Press.
- [38] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:arXiv:1803.07640
- [39] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.
- [40] Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. 2004. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE symposium on information visualization*. Ieee, 17–24.
- [41] Wei Gong, Ee-Peng Lim, and Feida Zhu. 2015. Characterizing silent users in social media communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.
- [42] Google. 2009. Google dashboard. (2009). <https://myaccount.google.com/dashboard>
- [43] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. <https://doi.org/10.5281/zenodo.4461265>
- [44] Alon Y Halevy. 2019. The Ubiquity of Subjectivity. *IEEE Data Eng. Bull.* 42, 1 (2019), 6–9.
- [45] Wu He, Xin Tian, Ran Tao, Weidong Zhang, Gongjun Yan, and V. Akula. 2017. Application of social media analytics: a case of analyzing online hotel reviews. *Online Inf. Rev.* 41 (2017), 921–935.
- [46] Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Proceedings of the CSCW 2021 Workshop—Investigating and Mitigating Biases in Crowdsourced Data. *arXiv preprint arXiv:2111.14322* (2021).
- [47] Walt Hickey. 2015. Be suspicious of online movie ratings, especially Fandango’s. *FiveThirtyEight*, Available at: <http://fivethirtyeight.com/features/fandango-movies-ratings> (2015).
- [48] Nan Hu, Paul A. Pavlou, and Jie (Jennifer) Zhang. 2009. Overcoming the J-Shaped Distribution of Product Reviews, Vol. 52. Communications of the ACM.

- [49] Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.
- [50] Milena Janic, Jan Pieter Wijbenga, and Thijs Veugen. 2013. Transparency enhancing tools (TETs): an overview. In *2013 Third Workshop on Socio-Technical Aspects in Security and Trust*. IEEE, 18–25.
- [51] Youngseung Jeon, Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. ChamberBreaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.
- [52] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 781–789.
- [53] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
- [54] Elahe Kani-Zabihi and Martin Helmholz. 2011. Increasing service users' privacy awareness by introducing on-line interactive privacy features. In *Nordic Conference on Secure IT Systems*. Springer, 131–148.
- [55] Hülya Karaman. 0. Online Review Solicitations Reduce Extremity Bias in Online Review Distributions and Increase Their Representativeness. *Management Science* 0, 0 (0), null. <https://doi.org/10.1287/mnsc.2020.3758>
- [56] Jan Kolter, Michael Netter, and Günther Pernul. 2010. Visualizing past personal data disclosures. In *2010 International Conference on Availability, Reliability and Security*. IEEE, 131–139.
- [57] Po-Ming Law, Leo Yu-Ho Lo, Alex Endert, John Stasko, and Huamin Qu. 2021. Causal Perception in Question-Answering Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 603, 15 pages. <https://doi.org/10.1145/3411764.3445444>
- [58] Angela Siew Hoong Lee, Ka Leong Daniel Chong, and Nicholas Chan Khin Whai. 2019. OpinionSeer: Text Visualization on Hotel Customer Reviews of Services and Physical Environment. In *Information Science and Applications 2018*, Kuinam J. Kim and Nakhoon Baek (Eds.). Springer Singapore, Singapore, 337–349.
- [59] Xinxin Li and Lorin M Hitt. 2008. Self-selection and information role of online product reviews. *Information Systems Research* 19, 4 (2008), 456–474.
- [60] Yuliang Li, Aaron Xixuan Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. 2019. Subjective databases. *arXiv preprint arXiv:1902.09661* (2019).
- [61] Sunghoon Lim and Conrad S Tucker. 2017. Mitigating online product rating biases through the discovery of optimistic, pessimistic, and realistic reviewers. *Journal of Mechanical Design* 139, 11 (2017).
- [62] Pei-Jung Lin, Eleri Jones, and Sheena Westwood. 2009. Perceived risk and risk-relievers in online travel purchase intentions. *Journal of Hospitality Marketing & Management* 18, 8 (2009), 782–810.
- [63] Yinhai Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [64] Dina Mayzlin, Yaniv Dover, and Judith Chevalier. 2014. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review* 104, 8 (August 2014), 2421–55. <https://doi.org/10.1257/aer.104.8.2421>
- [65] Dina Mayzlin, Yaniv Dover, and Judith Chevalier. 2014. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review* 104, 8 (August 2014), 2421–55. <https://doi.org/10.1257/aer.104.8.2421>
- [66] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* (2010), 1–1.
- [67] Mozilla. 2013. Lightbeam add-on for Firefox. (2013). <https://www.mozilla.org/en-US/lightbeam/>
- [68] Tamara Munzner. 2014. *Visualization analysis and design*. CRC press.
- [69] Rosie Murphy. 2020. Local Customer Review Survey. *Bright Ideas/Research* (12 2020). <https://www.brightlocal.com/research/local-consumer-review-survey/>
- [70] Rosanna Nagtegaal, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. 2020. Designing to Debias: Measuring and Reducing Public Managers' Anchoring Bias. *Public Administration Review* 80, 4 (2020), 565–576.
- [71] Rosanna Nagtegaal, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. 2020. Designing to Debias: Measuring and Reducing Public Managers' Anchoring Bias. *Public Administration Review* 80, 4 (2020), 565–576.
- [72] 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* (2021).
- [73] Brian A Nosek, Anthony G Greenwald, and Mahzarin R Banaji. 2005. Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin* 31, 2 (2005), 166–180.
- [74] Haekyu Park, Hyunsik Jeon, Junghwan Kim, Beunguk Ahn, and U Kang. 2017. Uniwalk: Explainable and accurate recommendation for rating and network data. *arXiv preprint arXiv:1710.07134* (2017).
- [75] Evan M Peck, Sofia E Ayuso, and Omar El-Etr. 2019. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [76] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. <https://doi.org/10.1016/j.jesp.2017.03.003>

[/doi.org/10.1016/j.jesp.2017.01.006](https://doi.org/10.1016/j.jesp.2017.01.006)

- [77] Emilee Rader. 2014. Awareness of behavioral tracking and information privacy concern in facebook and google. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, 51–67.
- [78] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3173574.3173677>
- [79] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.
- [80] Verena Schoenmueller, Oded Netzer, and Florian Stahl. 2018. The Extreme Distribution of Online Reviews: Prevalence, Drivers and Implications. *SSRN Electronic Journal* (01 2018). <https://doi.org/10.2139/ssrn.3100217>
- [81] Verena Schoenmüller, Oded Netzer, and Florian Stahl. 2019. The extreme distribution of online reviews: Prevalence, drivers and implications. *Columbia Business School Research Paper* 18-10 (2019).
- [82] Riyaz Sikora and Kriti Chauhan. 2011. Estimating sequential bias in online reviews: A Kalman filtering approach. *Knowledge Based Systems - KBS* 27 (01 2011). <https://doi.org/10.1016/j.knosys.2011.10.011>
- [83] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.
- [84] Anumeha Srivastava. 2016. Awareness Surveys: The Data-Driven Way to Read People's Minds. *Human of Data* (2016).
- [85] Jonathan AC Sterne, Matthias Egger, and David Moher. 2008. Addressing reporting biases. *Cochrane handbook for systematic reviews of interventions: Cochrane book series* (2008), 297–333.
- [86] Jennifer L. Stevens, Brian I. Spaid, Michael Breazeale, and Carol L. Esmark Jones. 2018. Timeliness, transparency, and trust: A framework for managing online customer complaints. *Business Horizons* 61, 3 (2018), 375–384. <https://doi.org/10.1016/j.bushor.2018.01.007>
- [87] Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A Simple Framework for Opinion Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5789–5798. <https://doi.org/10.18653/v1/2020.acl-main.513>
- [88] Jacob Thebault-Spieker, Daniel Kluver, Maximilian A Klein, Aaron Halfaker, Brent Hecht, Loren Terveen, and Joseph A Konstan. 2017. Simulation experiments on (the absence of) ratings bias in reputation systems. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–25.
- [89] Chih-Fong Tsai, Kuanchin Chen, Ya-Han Hu, and Wei-Kai Chen. 2020. Improving text summarization of online hotel reviews with review helpfulness and sentiment. *Tourism Management* 80 (2020), 104122. <https://doi.org/10.1016/j.tourman.2020.104122>
- [90] C. Walker and Scott Buttinger. 2017. Towards Mitigating Bias in Online Reviews : An Application to Amazon.
- [91] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Padén, 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.
- [92] Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246.
- [93] Xiaolan Wang, Yoshihiko Suhara, Natalie Nuno, Yuliang Li, Jinfeng Li, Nofar Carmeli, Stefanos Angelidis, Eser Kandogan, and Wang-Chiew Tan. 2020. ExtremeReader: An Interactive Explorer for Customizable and Explainable Review Summarization (WWW '20). Association for Computing Machinery, New York, NY, USA, 176–180. <https://doi.org/10.1145/3366424.3383535>
- [94] Ding Wu, Xunhua Guo, and Guoqing Chen. 2017. Mitigating the Dependence Bias in Online Ratings: A “Consider-the-Opposite” Strategy for Scale Prompting. (2017).
- [95] Ding Wu, Xunhua Guo, and Guoqing Chen. 2017. Mitigating the Dependence Bias in Online Ratings: A “Consider-the-Opposite” Strategy for Scale Prompting. (2017).
- [96] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. 2010. OpinionSeer: Interactive Visualization of Hotel Customer Feedback. *IEEE Transactions on Visualization and Computer Graphics* (November 2010), 1109–1118. <https://www.microsoft.com/en-us/research/publication/opinionseer-interactive-visualization-hotel-customer-feedback/>
- [97] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv.* 50, 2 (2017). <https://doi.org/10.1145/3057270>
- [98] Markus Zanker and Daniel Ninaus. 2010. Knowledgeable explanations for recommender systems. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 657–660.

- [99] Angeliki Zavou, Vasilis Pappas, Vasileios P Kemerlis, Michalis Polychronakis, Georgios Portokalidis, and Angelos D Keromytis. 2013. Cloudopsy: An autopsy of data flows in the cloud. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, 366–375.
- [100] Xiong Zhang, Jonathan Engel, Sara Evensen, Yuliang Li, Çağatay Demiralp, and Wang-Chiew Tan. 2020. Teddy: A System for Interactive Review Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [101] Xiaoying Zhang, Hong Xie, Junzhou Zhao, and John CS Lui. 2019. Understanding assimilation-contrast effects in online rating systems: modelling, debiasing, and applications. *ACM Transactions on Information Systems (TOIS)* 38, 1 (2019), 1–25.
- [102] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval* 14, 1 (2020), 1–101.
- [103] Tianxiang Zheng, Feiran Wu, Rob Law, Qihang Qiu, and Rong Wu. 2021. Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach. *International Journal of Hospitality Management* 92 (2021), 102658.
- [104] Tianxiang Zheng, Feiran Wu, Rob Law, Qihang Qiu, and Rong Wu. 2021. Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach. *International Journal of Hospitality Management* 92 (2021), 102658. <https://doi.org/10.1016/j.ijhm.2020.102658>

Received January 2022; revised April 2022; accepted August 2022