Informing Users about Data Imputation: Exploring the Design Space for Dealing with Non-Responses

Ananya Bhattacharjee[1], Haochen Song[1], Xuening Wu[1], Justice Tomlinson[1], Mohi Reza[1], Akmar Ehsan Chowdhury[1], Nina Deliu[2,3], Thomas W. Price[4], Joseph Jay Williams[1]

[1]Intelligent Adaptive Interventions Lab, University of Toronto
[2]MEMOTEF Department, Sapienza University of Rome
[3] MRC Biostatistics Unit, Cambridge University
[4]Department of Computer Science, North Carolina State University
ananya@cs.toronto.edu, fred.song@mail.utoronto.ca, shirleyxuening.wu@mail.utoronto.ca, justice.tomlinson@mail.utoronto.ca, mohireza@cs.toronto.edu, akmar.chowdhury@mail.utoronto.ca, nina.deliu@uniroma1.it, twprice@ncsu.edu, williams@cs.toronto.edu

Abstract

Machine learning algorithms often require quantitative ratings from users to effectively predict helpful content. When these ratings are unavailable, systems make implicit assumptions or imputations to fill in the missing information; however, users are generally kept unaware of these processes. In our work, we explore ways of informing the users about system imputations and experiment with imputed ratings and various explanations required by users to correct imputations. We investigate these approaches through the deployment of a text messaging probe to 26 participants to help them manage their psychological wellbeing. We provide quantitative results to report users' reactions to correct vs. incorrect imputations and the potential risks of biasing their ratings. Using semi-structured interviews with participants, we characterize the potential trade-offs regarding user autonomy and draw insights about alternative ways of involving users in the imputation process. Our findings provide useful directions for future research on communicating imputation and interpreting user non-responses.

Introduction

Many systems rely on human feedback for improvement (Ricciardelli and Biswas 2019; Khan et al. 2021). For example, a text messaging system that aims to help people manage their stress may ask a user to rate a supportive text they received in terms of usefulness (Figueroa et al. 2022). The ratings are usually quantitative; for example, they could be obtained on a scale of 1 to 5, where 1 indicates "not useful at all," and 5 indicates "very useful". Eventually, the rating data is fed to machine learning algorithms to decide which messages might be the most beneficial to future users. Among other statistical and technical aspects, the continuous improvement of such systems relies heavily on the quantity and quality of the collected data. Removing even a single user's response implies that past users' data matters more, although the user may not have responded because they saw the message as good and therefore saw no reason to comment, or thought it was bad and did not want to spend any effort. When users become non-responsive, the loss of data reduces the algorithms' capability to deliver optimized service to users (Figueroa et al. 2021).

Typically, the problem of missing data is either ignored with a complete case analysis (Figueroa et al. 2022; White and Carlin 2010) or addressed by designing complex machine learning algorithms to impute missing data (Jerez et al. 2010). Imputation refers to a set of techniques to replace the missing values in a dataset with substituted values that can retain most of the information about the dataset (Jerez et al. 2010; Richman, Trafalis, and Adrianto 2009; Bhattacharjee, Bayzid et al. 2020). However, imputation techniques fill missing data with estimates, typically without user involvement. The algorithms run in the system's backend, taking no user feedback on the accuracy of imputations. Their performance varies based on several factors, such as the amount of missing data, dataset linearity, and hyperparameter choice (Du, Hu, and Zhang 2020). Additionally, these techniques often work well on specific datasets but falter when the dataset changes (Seaman, Bartlett, and White 2012).

In contrast to these works, we investigate the feasibility of involving users in the imputation process to gather more data and make the process more transparent. We design a framework taking motivation from the literature on machine learning transparency (Rader, Cotter, and Cho 2018) and scrutability (Mahmoud 2021; Pardos, Fan, and Jiang 2019); when users do not respond with evaluation ratings of a message, they are directly informed of the value a system might impute, and we let them correct it if they feel necessary, e.g., if they disagree with the proposed value. This framework presents several research opportunities, such as the possibility of reducing the amount of missing data through enhanced user participation or examining user responses to both correct and incorrect imputed values. Additionally, we can investigate the design of messages to effectively communicate the importance of providing ratings to users.

Our research explores the following novel idea: *To what extent can we build a transparent framework for the missing data imputation process (at least partially) that allows users to see the imputation and correct it if necessary?* To explore the feasibility of this idea, we investigated the following two research questions:

- **RQ1:** How does such a framework affect different components of users' response behavior (e.g., response rate, effort in generating ratings)?
- **RQ2:** What potential challenges might arise when users interact with the framework and correct the imputed value?

Our investigation into the design of our framework started with interviews with four professionals with expertise in machine learning, statistics, and human-computer interaction. Based on their responses and a review of the literature, we identified several design elements, related to user involvement and various ways of explaining the imputation process. To explore the proposed framework, we ran a longitudinal study with 26 participants over a period of 9 days. The study involved a text messaging probe that provided suggestions for managing stress and promoting psychological wellbeing. During the whole study period, after an initial rating prompt, non-respondents were randomly assigned to receive a regular reminder to provide a rating or a version of our proposed messages which informed users of imputations and allowed them to correct them. We integrated all design elements from our interviews with professionals into designing the messages; we experimented with different explanations of the imputation process and explained why users should give their feedback. Our results showed that the proposed messages were comparable to regular reminders in terms of eliciting user responses. Participants found it easier to rate messages when they were given the imputed ratings as a benchmark, as opposed to just being reminded to rate the messages with no imputed ratings. This highlights the potential of involving users in the imputation process to make the rating generation process less burdensome for users. However, we observed more responses than expected for the cases where people's given ratings were the same as the imputed values. Our findings also indicate that people tended to give slightly higher ratings than the imputations they saw on average. Lastly, we provide important qualitative feedback regarding various elements of user experience (e.g., participants may feel they have less control over their interaction with the proposed framework, explaining the imputation process should involve mentioning examples and sources of data).

Our contributions include the following:

- The proposal of a novel framework that involves users in the data imputation process
- Insights about different components of interacting with imputed ratings and providing user feedback (e.g., response rate, effort in generating ratings) from the deployment of a text messaging probe
- A set of considerations for future work, involving potential user bias and alternative prototypes of communicating data imputation

Related Work

In this section, we first provide a review of the user non-response issue, specifically focusing on the key reasons behind it. Then we go on to explain how one can address this issue by taking motivation from the literature on machine learning transparency and scrutability.

User Non-Response

User non-response, defined as either an agent's failure to reply or an incomplete reply, presents significant challenges in data analysis (MacDonald et al. 2009). It reduces the sample size, thereby negatively impacting statistical power and precision. In addition, if neglected, it can carry different degrees of bias depending on the mechanism driving non-responses (Rogelberg and Stanton 2007).

Several psychology and user experience (UX) theories propose that users' response behaviors are primarily influenced by two factors: (1) a subjective lack of motivation and (2) the objective presence of barriers (Ajzen 2005; Davis 1989). In the context of user non-response, the perceived importance of reply links to users' perceived ease of generating and sending a response, and both factors determine whether users are motivated to reply to the message. Similarly, the coupled notions, pragmatic quality, and hedonic quality (Hassenzahl 2008), decompose the potential reasons behind user non-response. Pragmatic quality focuses on the usability aspects which relate to the barriers or burdens associated with responses, while hedonic quality focuses on the enjoyment aspects which determine whether users will be motivated to reply or not. Further, according to Fogg behavioral model (Fogg 2009), behavior is the result of motivation, ability (or simplicity of the task), and prompts. This model suggests that besides demotivation and inability, there is a third limiting factor: whether people are prompted to a behavior. For a prompt to be effective, it should be related to the expected behavior and take place when people have sufficient motivation and ability (Fogg 2009; Chounta and Nolte 2022). With the potential to fulfill these criteria, we posit that communicating the assumptions (imputed ratings) can act as a cue to prompt users for ratings.

Previous research has unveiled a significant link between the absence of motivation and user non-response (Groves, Presser, and Dipko 2004; Chou et al. 2022). Studies have shown that response rates increase significantly when participants are interested in the study topic (Law et al. 2016). Furthermore, response facilitation approaches, such as establishing the importance of the survey, providing incentives, and enhancing question relevance, have been proven effective in motivating users to respond (Rogelberg and Stanton 2007; Halbesleben and Whitman 2013; MacDonald et al. 2009). In addition to demotivation, the cognitive load and mental effort required for generating and dispatching a response also significantly contribute to user non-response (MacDonald et al. 2009). The Lazy User Theory asserts that users tend to opt for the path of least resistance when given multiple options (Tétard and Collan 2009). This theory underscores the notion that even though the cognitive load associated with generating and sending a rating might not be prohibitive, it can still influence the response rate if the process is not streamlined to be as effortless as possible. Therefore, reducing barriers and simplifying the task can be instrumental in encouraging users to leave ratings.

In this paper, we are interested in knowing how transparency (Rader, Cotter, and Cho 2018) and scrutability (Mahmoud 2021; Pardos, Fan, and Jiang 2019) affect the user non-response issue; we intend to observe whether our proposed framework reduces (or increases) users' demotivation and the associated cognitive load.

Transparency and Scrutability

Transparency of a system enables a user to understand how it works (Rader, Cotter, and Cho 2018; Ehsan et al. 2021). The importance of transparency has already been recognized in expert systems (e.g., clinical applications, criminal justice) (Jaffe et al. 2006; Montes and Luna 2021). Stakeholders often express distrust and discomfort when they do not understand how the black box artificially intelligent models are making their decisions (Amershi et al. 2019; Yang et al. 2020; Krause, Perer, and Ng 2016). However, researchers suggest providing informed justification (e.g., an individual is consistently getting messages to do physical activities because the system thinks they are overweight) would help users follow and trust the decisions made by those models (Rader, Cotter, and Cho 2018; Kim et al. 2022; Woźniak et al. 2020; Krause, Perer, and Ng 2016; Liao et al. 2022; Cai, Jongejan, and Holbrook 2019).

The concept of transparency can be extended by scrutability (Mahmoud 2021; Pardos, Fan, and Jiang 2019; Tintarev 2007) – the quality of a system that allows a user to first understand its functioning (e.g., how it is providing a recommendation) and then enables the user to correct its assumptions when necessary. For example, if a system imputes that the user would like an intervention, whereas in reality the user does not, scrutability ensures that the user will be able to point out the flaws in the system's decision-making process. Eventually, it allows systems to make iterative improvements and provide better recommendations in the future (Doan 2018; Tintarev 2007).

In many ways, scrutability has the potential to address the two key reasons we mentioned for user non-response. First, scrutability potentially mitigates user non-response by addressing motivation. Social Determination Theory (SDT) suggests that autonomy, competence, and relatedness, the three basic psychological needs, fuel intrinsic motivation (Ryan and Deci 2000). A scrutable system has the potential to provide autonomy through user choices and corrections, relatedness by including users in the rating generation process, and competence via transparency (Cai, Jongejan, and Holbrook 2019). By fulfilling these "be-goals" as per Hassenzahl (2008), the perceived hedonic quality is increased, which encourages users to interact intrinsically.

Second, scrutability may reduce users' burden by showing them the algorithm-generated rating and letting them make corrections. Compared with directly asking for a rating where users need to retrieve and evaluate different aspects of their experiences, we posit that showing them an imputed rating can reduce the cognitive load by altering the salience of different pieces of information, which highlights the most important and relevant information for users to generate a rating (Shah and Oppenheimer 2008). In other words, communicating the generated imputation can play a role in information filtering that reduces cognitive load by reducing information overload (Quiroga, Crosby, and Iding 2004; Shah and Oppenheimer 2008).

In this paper, we see the quality of scrutability as a means for collecting data as well. With careful design, we believe a framework with scrutability can motivate users to provide information about their preferred content with less burden.

Semi-Structured Interviews with Professionals

Given the novel and unexplored framework we are proposing in this work, we were first concerned with getting professionals' perspectives and determining the fundamental design aspects, such as the type of interventions or the choice of wording, that may be worth exploring in a subsequent field experiment. We leveraged existing literature and semi-structured interviews with four professionals to explore various opportunities and trade-offs of sharing imputed ratings with users.

Participants

The interviewees were recruited from email invitations and word-of-mouth. All of them had post-graduate degrees – one was a faculty member (Pr1), two were graduate students (Pr2 and Pr3), and the other was working at a software company (Pr4). All of the participants have expertise in machine learning and statistics and have worked on building digital mental health interventions. Besides, these participants also had experience publishing papers at HCI conferences. Their mean age was 31.3 ± 3.2 years old; they were of two genders (2 women, 2 men; other options were offered) and two ethnicities (2 Asian, 2 White). At the time of the study, they were living in North America.

Procedure

In the individual interviews, we first explained the motivation for our study to professionals. Then we wanted to gather their feedback on the feasibility of the approach (i.e., what if systems communicated their imputed ratings to users?). Our questions included, but were not limited to, 'What do you think of our approach?', and 'Why do you think this approach would work? Why wouldn't it work?'. To understand the design space, we also asked questions like 'What is something that can be added to make this idea even better?' and 'What kinds of instructions might we provide to explain what the system is doing, how the data is being used, and why the assumption is being made?'. The semi-structured nature of the interviews allowed us to deviate from the interview script and ask clarifying questions whenever necessary.

The interviews were conducted via the Zoom video conferencing platform. One member of the research team conducted the interviews. Each interview lasted 30–60 minutes, and interviewees were compensated USD $12 for their time.

Data Analysis

After transcribing and cleaning the interview data, we followed a thematic approach to analyze the qualitative data (Cooper et al. 2012). Two members of the research team (referred to as "coders") first reviewed all transcripts to become familiar with the data. The data segments were then assigned to distinct codes using the open-coding process (Khandkar 2009). Each coder developed a preliminary codebook independently first; then, they proceeded to create a shared codebook after meeting several times when they identified overlapping codes and refined code definitions. The shared codebook was applied to a subset of the data, and coders again met to refine the codebook. After reaching a consensus through this iterative process, the coders applied the final codebook to separate halves of the data.

Findings

Below we highlight the important design elements that came up from our interviews and review of the literature.

User involvement An important motivation of our framework is to involve users in the imputation process, by showing them the imputed rating and allowing them to change if necessary. Literature indicates that soliciting user input can increase engagement and foster a sense of individualized experience and control over the interaction (Stout, Villegas, and Kim 2001; Segijn et al. 2021; Bansal et al. 2019).

However, excessive demands for user input could render the framework overly burdensome (Suh et al. 2016). Pr1 also anticipated this issue:

> "*Many times, people don't rate, simply because they are busy. So, if they have to still respond to you with your new design, you will still get a lot of non-response. ... You can reduce users' effort by giving them a number and allow them to change if it's wrong.*"

At the same time, Pr3 and Pr4 emphasized that an imputed rating should not be treated as a user rating without their explicit confirmation. Pr4 said:

> "*Let's say the system is pretty good in the beginning. It kind of can assume the rating that is close to my heart. That would, I think, lose the novelty. In that case, after two or three days I would start to see, hey the message always gives my real rating, so why bother looking at it? After that even if the ratings are wrong, I would not probably notice.*"

These comments motivated us to come up with a framework, where we communicate the imputed rating to users, and ask them to respond no matter whether they agree or disagree with the imputation. If they agree, we ask them to respond with "Yes", otherwise with their actual rating.

Explaining the importance of imputation and user rating The professionals also anticipated that communicating the imputations to people without any relevant information might come across as confusing and overwhelming. All of them emphasized that users should be made aware of the importance of their feedback. They suggested that the value of user feedback should be made explicit to users. Pr1 said,

> "*People are not gonna read the messages like how you read it. They may not know why their feedback is important. ... The system should say something like "Your responses help us give you better texts, so please respond to make the system more useful.*"

We incorporate this suggestion by adding a sentence along the lines of the following text in the deployment: *"We need your rating so that we can give you and others the most useful possible messages."*

Moreover, Pr2 expressed that we should also explain why the framework is showing them the imputed rating. Literature suggests that context-relevant cues can reduce the difficulty of retrieving and integrating information (Shah and Oppenheimer 2008). In this case, the imputed ratings can act as the cues that make the rating-relevant information more salient and accessible for users, thereby reducing user burden.

We incorporate this suggestion by adding a sentence along the lines of the following text:

"We will make an assumption about your rating to save time and reduce your burden."

Explaining imputation process Our interviewees suspected that a regular user may not be aware that when they do not provide a rating, their missing rating may be imputed at the back end. Even those who know, may not be familiar with the imputation process. However, prior research suggests that communicating how recommendations are made can help users become more engaged. It can also reduce frustration, in case the provided recommendation is not satisfactory, or in our case an incorrect imputation (Zhou and Chen 2018; Alvarez-Melis et al. 2021).

Our interviewees suggested experimenting with various framings about the process of imputation. Pr1 and Pr2 reminded us that recommendations can be made using data from one particular user's past behavior or how other people have reacted to a particular product. Again, in cases where there is limited data available, sophisticated algorithms may resort to random imputation. Hence, we experimented with three framings:

- Telling users that the imputed rating is a random number (Kalton and Kish 1984)
- Telling users that the imputed rating has been generated from their past interaction with the system (Hawthorne, Hawthorne, and Elliott 2005)
- Telling users that the imputed rating has been generated from other users' past interaction with the system (Andridge and Little 2010)

Choice of wording and sentences We also asked our participants about how we should convey the imputed ratings. Possible choices of words and phrases included *assume, predict, our best guess,* and *impute*. Professionals generally supported the use of the word *assume* or *assumption* because the word did not have technical connotation (unlike *impute*), could communicate that the ratings could be changed (unlike *impute* or *predict*), and at the same time formal (unlike *our best guess*). Hence, we showed a sentence along the lines of the following in the text message deployment:

We will assume you rate this message a [[imputed rating]].

For the rest of the paper, we use the words *assume/assumption* and *impute/imputation* interchangeably, unless mentioned otherwise.

Deployment of a Text Messaging Probe

In this phase, we conducted a longitudinal study using a text messaging probe designed to aid individuals in managing stress and psychological wellbeing. Given past experiences with missing data and user non-response in digital mental health tools (Bhattacharjee et al. 2023a), the research team applied their proposed framework in a similar setting. Text messaging was chosen for its accessibility and ubiquity in promoting healthy behaviors, and the study also offered the opportunity to test our framework in real-life contexts (Feroz et al. 2019; Bhattacharjee et al. 2023b).

Participants

We recruited 26 participants from a large introductory programming course at a major North American University. We refer to them as P1–P26. Participants were recruited through email invitations, and they did not have to fulfill any inclusion criteria. Their mean age was 20.3±0.3 years old. They were of two genders (20 women, 6 men; other options were offered) and multiple ethnicities (18 Asian/Pacific Islander, 6 White, 1 Black, 1 undisclosed).

Study Procedure

Formation of assumption and reminder messages We integrated all of the design elements we identified with the messages to communicate imputations. Our messages contained the texts "please let us know, so we can give you and others the most useful possible messages" to communicate the importance of giving ratings and "We will make an assumption about your rating to save time and reduce your burden" to explain the purpose of making assumptions. We made the choice to use the word 'assume' as an alternative to 'impute'. We refer to these messages as *assumption messages*.

Assumption messages could contain any of the following three texts to explain the imputation process. We created the following three explanations.

- **Assumption 1**: Many technologies make these assumptions randomly.
- **Assumption 2**: Many technologies make these assumptions based on a particular user's past interaction with the technology.
- **Assumption 3**: Many technologies make these assumptions based on other users' interactions with the technology.

While there exists a range of imputation techniques, we were interested in observing people's reactions to the whole space of imputed ratings - how do people's reactions vary when the alignment between the imputed rating and assumed rating ranges from identical to completely opposite? Hence, we evenly spaced the imputed ratings from 1 to 5.

We decided to use **regular reminders** (that do not show any imputed ratings) as a baseline for our experiment, using which we could compare the reactions to our assumption messages. However, the regular reminders also contained similar texts to communicate the importance of giving ratings.

For this study, we use the term *follow-up message* to refer to assumption and regular reminder messages together, since both of them were sent as a follow-up when participants did not respond. Table 1 shows all the different follow-up messages we designed.

Formation of supportive messages to help participants manage stress The supportive text messages for helping people manage stress were inspired by theories of cognitive behavioral therapy (CBT) (Rothbaum et al. 2000) and targeted to help people manage their stress and accompanying negative emotions. Our research team, which included faculty members and graduate students of human-computer interaction and psychology, iteratively developed a message bank containing 15 messages. The messages were inspired by prior work, online resources, and books (Haarhoff and Thwaites 2015; Kaye 2017; Niemiec 2013a,b; Robotham and Julian 2006).

Pilot Study to Inform the Design of the Main Deployment Before starting the main deployment of the text messaging probe, we conducted a pilot study with 5 participants to decide on the time interval between various messages and refine our designed messages. These people (who were not part of the main deployment) were recruited in the same way from the same introductory programming course. We recruited them for 9 days and sent messages at different time intervals. After their interaction with the messages for 9 days, these participants suggested that a user should be allowed an hour to process and reflect on the text on stress management. After they were asked for ratings, the text messaging probe should also wait for some time before showing the imputed rating, as some participants might be motivated to give ratings without any follow-up. However, the wait time should not be too long to make people forget about their reactions to seeing the original supportive message. The suggestions for ideal time ranged from 30-60 minutes. They also suggested that the sentence containing the imputed rating should be highlighted so that the users could easily notice it.

Design of the Main Deployment Based on the suggestions made by the pilot study participants and the already identified design elements, we designed the deployment work. The deployment study ran for 9 days. On each day, participants received a text message at a random time between 6 a.m. and midnight in their timezone. The message was selected randomly from the message bank we constructed. It was also ensured that participants did not receive the same message more than once during the study period.

After one hour of receiving the message, participants were asked to rate the message on a 1 to 5 scale. We sent the following prompt asking for ratings (Rating message):

> "*How glad are you receiving that message at the moment you did, based on the state you were in? Please*

Assumption 1	Hello, [[user]]! We noticed you didn't get around to replying to our previous prompt yet, which is completely fine. However, if you are unable to respond today, we will make an assumption about your rating to save time and reduce your burden. **WE WILL ASSUME YOU RATE TODAY'S MESSAGE A [[imputed rating]]**. *Many technologies make these assumptions randomly.* Please type "yes" if you agree with our assumption, or text back your actual rating, so that we can give you and others the most useful possible messages.
Assumption 2	Hello, [[user]]! We noticed you didn't get around to replying to our previous prompt yet, which is completely fine. However, if you are unable to respond today, we will make an assumption about your rating to save time and reduce your burden. **WE WILL ASSUME YOU RATE TODAY'S MESSAGE A [[imputed rating]]**. *Many technologies make these assumptions based on a particular user's past interaction with the technology.* Please type "yes" if you agree with our assumption, or text back your actual rating, so that we can give you and others the most useful possible messages.
Assumption 3	Hello, [[user]]! We noticed you didn't get around to replying to our previous prompt yet, which is completely fine. However, if you are unable to respond today, we will make an assumption about your rating to save time and reduce your burden. **WE WILL ASSUME YOU RATE TODAY'S MESSAGE A [[imputed rating]]**. *Many technologies make these assumptions based on other users' interactions with the technology.* Please type "yes" if you agree with our assumption, or text back your actual rating, so that we can give you and others the most useful possible messages.
Regular Reminder	Hello, [[user]]! We're just sending a friendly reminder to encourage you to reply to the previous prompt with your rating. We value your feedback as it helps us send more useful messages to you and other users.

Table 1: Follow-up messages that were sent through text messages. The italicized sentences show differences across three assumption messages.

reply with a rating between 1 (Not glad at all) to 3 (neutral) to 5 (super glad to receive it)?"

If the participant provided a rating within 45 minutes, no follow-up message was sent, and the conversation for the day would be over. However, if the participant did not, they would receive a follow-up message, which could be either one of the three assumption messages or the regular reminder message. They were scheduled to receive assumption messages on a random subset of 7 days, while the rest two days were assigned for reminders. We also note that scheduling a follow-up message for a particular day does not necessarily mean they would see that message on that day; if a participant responded with their rating within 45 minutes of the initial rating prompt, we did not send them the follow-up message. Figure 1(a) illustrates the messaging protocol for each day during the study period, and 1(b) shows how participants were scheduled to receive follow-up messages during the study period.

Participants were sent messages using Twilio, an automated message delivery platform. They were informed beforehand that their level of interaction with the text messages would not affect their compensation. After completing the study, they were invited to take part in a follow-up semi-structured interview to provide their feedback on the assumption messages. Ten participants agreed to take part in the interviews. Our interview questions included, but were not limited to:

- What did you like or not like about the experience of receiving assumption messages?
- What kinds of instructions might we provide to explain what the framework is doing, or why the assumption is being made?
- What were the advantages or disadvantages of receiving assumption messages over regular reminders that do not assume your rating?
- How can we further improve our current prototype messages?

The interviews were conducted via Zoom videoconferencing platform and lasted 15-40 minutes. All interviewees were compensated USD $12 for their participation.

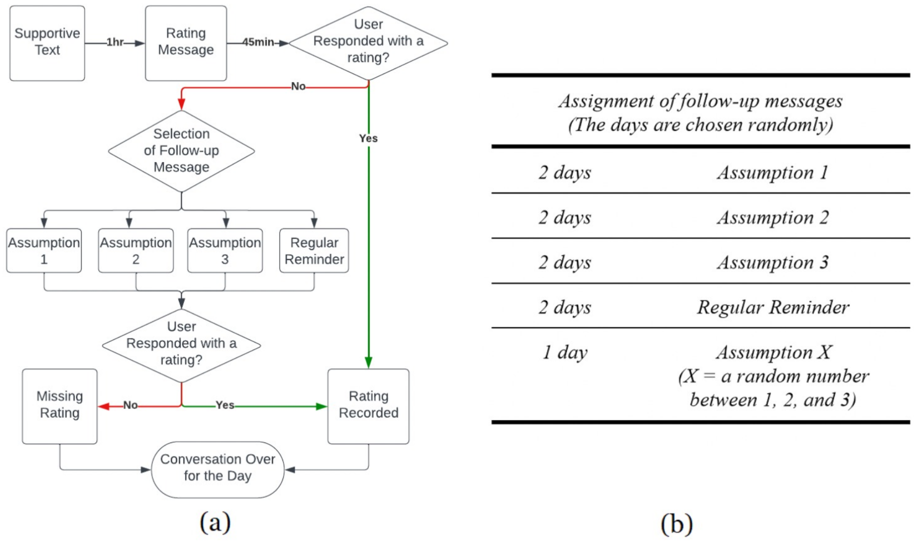
Data Analysis

We calculated the response rates of both assumption and reminder messages. Specifically, if a participant provided a rating in response to a reminder or assumption message, we counted that as a response. We also analyzed the differences between the assumed ratings and the actual ratings people gave. The value $\alpha = 0.05$ is used as the significance level for testing the null hypothesis of no difference.

The interview data were analyzed in the same way as described in the previous interviews with professionals.

Ethical Considerations

The study was approved by the Research Ethics Board at the University of Toronto, Canada. Additionally, we are fully aware that research involving psychological wellbeing raises a few ethical issues that we carefully considered throughout the study (Kornfield et al. 2022). We informed participants at the beginning of the study that ours was not a crisis service, although provided them with the contact information of crisis services. Participants had the option to leave the study at any stage. We reviewed all text responses by participants on a daily basis to notice whether there was any indication of self-harm or suicidal ideation. In the unlikely



Assignment of follow-up messages (The days are chosen randomly)	
2 days	*Assumption 1*
2 days	*Assumption 2*
2 days	*Assumption 3*
2 days	*Regular Reminder*
1 day	*Assumption X* *(X = a random number between 1, 2, and 3)*

Figure 1: (a) Messaging protocol for each day during the study period (b) Schedule of follow-up messages during the study period

possibility that a message had such indication, research team members were trained to reach out to the sender and conduct Columbia-Suicide Risk Assessment protocol (Posner et al. 2008). Similar considerations were also applied to the interviews. Interviewees were informed that they could leave the conversation at any time or choose to not answer any question. However, no risk emerged during the study, and no follow-up assessment was conducted.

Findings

In this section, we initially explore the response rates to various messages, then present insights from user interviews regarding their interaction with assumption messages, and finally compare actual and imputed ratings to gain insights into users' response behavior.

Response rates before and after follow-up messages We had 26 participants, each of whom received 9 messages, resulting in $26 \times 9 = 234$ observations in total. The response rate was 41.45% (97/234) before any follow-up message was sent, while after follow-up messages, the response rate rose to 81.20% (190/234), showing an increase of 39.74% (93/234).

Recognizing that there was also a probability that some people might have responded more than 45 minutes after the original request for rating, we did not conduct any statistical tests to detect significant differences between the response rate before and after the follow-up messages. We left these findings as promising exploratory results.

Nevertheless, we provided comparisons between the assumption and regular reminder messages. As shown in Table 2, no statistically significant difference was detected in the response rates of all follow-up messages. Combining three assumption messages altogether, we see that they had generated slightly more responses than reminders on average (68% (73/107) vs. 67% (20/30)), but again with no statistical significance (P-Value = 0.88, χ^2 Statistic = 0.03).

User experience of interacting with follow-up messages Several themes, discussed below, emerged from our analysis of the semi-structured interviews with participants.

Comparing assumptions with regular reminders: People shared their different perspectives on interacting with the assumption messages and regular reminders. Many of them acknowledged that they had not encountered anything similar to assumption messages before, so initially, the assumption messages came across as 'confusing' for some. However, over the course of the study, they became accustomed to the framework communicating its assumptions to them. P2 expressed:

> "*[[I was bothered]] only the beginning, especially the first time and then the second time, I guess, and then I gradually got comfortable.*"

Participants pointed out several advantages of receiving assumption messages. For example, P15 mentioned that seeing an imputed rating helped them set a 'baseline' based on which they could compare their true ratings. In the case of regular reminders, they missed having that baseline, essentially making it more difficult for them to provide their true rating.

Feedback about autonomy and control: Participants had varying perspectives on the control they had over the interaction with the assumption messages. Some felt that the imputed ratings in the messages reduced their control, as *'somebody else was making a decision for me'* (P26). As a result, they preferred to have their own choice in determining whether they wanted to make a decision without seeing the imputed rating or not. On the other hand, some others did not encounter those feelings; they pointed out that although the system was providing its own assumption, it was

Response Rate				χ^2 Statistic	P-Value
Assumption 1	Assumption 2	Assumption 3	Regular Reminder		
73% (27/37)	67% (22/33)	65% (24/37)	67% (20/30)	0.65	0.89

Table 2: Comparisons between response rates in various conditions in different assumption and reminder messages. χ^2 statistics and P-value are calculated based on a Binomial setting ANOVA test, under the null hypothesis: $H_0 : p_{Assumption\,1} = p_{Assumption\,2} = p_{Assumption\,3} = p_{RegularReminder}$, with p being the response rate, and the alternative hypothesis H_1: at least one follow-up message's response rate is different from others'.

allowing the user to confirm whether the assumed rating was correct or not. This confirmation prompt, according to them, made them feel they had control over the final rating.

Feedback on wording and sentences: We received a range of feedback about our choice of the word "assume" or "assumption". People like P16 felt that the word came across as 'personal', while other words like 'predict' could remind people of the fact that there is a machine on the backend, taking away the 'human feeling' of the conversation. P15, on the other hand, felt 'assume' was a 'strong word' and opted for other phrases like 'our best guess' or 'we think you would say this'.

People also said that a one-sentence explanation of the process of generating assumed ratings could come across as ambiguous. P8, particularly focusing on Assumption 3, felt that it could be improved by adding specific examples or explanations. They commented:

> "*I think if you guys could have said that we're making this assumption based on the fact that 'You know, in the last year, this is what students felt', or 'during November or October, students tend to feel this way'. ... I think that would have helped me understand how are you coming up with these ratings.*"

Participants also gave us suggestions on how we could present these explanations differently. One common suggestion was instead of sending a single long text, we could have broken it down into two or three short texts and put important information in the first sentence. Another way of drawing attention was starting the beginning of a sentence with keywords (e.g., "Source").

Alternative ways to deliver assumption messages: Our interviewees suggested several alternative ways to deliver imputations as well. Many suggestions involved using multiple follow-up messages. P17 said that the first follow-up message could contain regular reminders without any assumptions, and if the participants still do not respond within a few hours, the assumption messages could come. This way, they felt, the framework could provide a longer window to users without the risk of influencing their ratings. Some participants also wanted to set up their own time window before the assumption messages came, as they felt the 45-minute time period in our study could be too long or too short, depending on a person's responsiveness, or the time a message is being sent.

Comparison between the user ratings and the imputed ratings The deployment of our proposed framework also allowed us to infer several insights about users' response behavior, particularly when we compared the imputed ratings people saw (Assumed Rating) and the ratings they provided (Actual Rating).

n	Mean Difference: M(Actual Rating - Assumed Rating)	Mean Absolute Difference: M(|Actual Rating - Assumed Rating|)
73	0.14±0.18	1.15±0.12

Table 3: Differences between actual ratings given by participants and the assumed ratings

Table 3 shows that people gave slightly higher ratings than the assumptions on average. For the cases where people provided a response to assumption messages ($n = 73$), the ratings provided by people were greater by 0.14 than the imputations on average.

We also looked into the relative frequency distribution of differences, to infer information about people's non-response. We divided differences between actual and assumed ratings into three groups: 1) no difference, when the participants provided the same rating as the imputed value or responded with 'Yes', 2) small difference, when the absolute difference between actual and assumed ratings was 1 or 2, and 3) large difference, when the absolute difference between actual and assumed ratings was 3 or 4.

A crucial factor in our decision to compare relative frequencies of absolute differences is the structure of the experiment. The absolute difference between actual and assumed ratings is more likely to be 0 or 1 rather than 3 or 4. This is because an absolute difference of 1 can occur when the actual rating is between 1 and 5, while an absolute difference of 4 can only occur when the actual rating is either 1 or 5. Hence, we calculated the expected relative frequencies for all values of absolute differences to make a comparison between expected and observed relative frequencies for each individual value of absolute difference between actual and assumed ratings.

Figure 2 shows the expected and observed relative frequencies for no, small, and large differences between actual and assumed ratings. Participants tended to respond 13% more than expected when the assumed ratings were the same as their actual ratings. People provided a similar number of responses as expected when the assumed ratings had a small difference from actual ratings; however, as the differences became large, the response rates tended to go about 50% lower than expected. Non-response to an assumption message, hence, might indicate that there is a high probability the assumed ratings were not close to users' actual ratings.

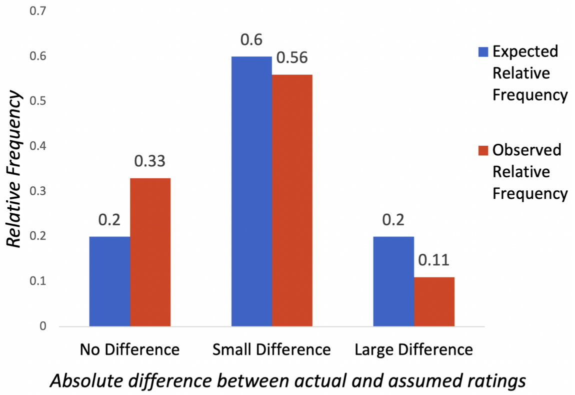


Figure 2: Expected and observed relative frequencies of absolute differences between actual and assumed ratings

Discussion

Our work contributes to the literature on transparency (Rader, Cotter, and Cho 2018) and scrutability (Mahmoud 2021) by exploring novel approaches to better involve users in the data collection and imputation process. Furthermore, it provides an alternative way of gathering more data that can support performing missing data imputation, such as by sharing imputed values with users and prompting for corrections, which can provide additional data. In this section, we first discuss the outcomes of our investigation and how they relate to our research questions. Then, we outline the limitations and future prospects of our work.

Impact on Response Behavior

RQ1: How does such a framework affect different components of users' response behavior (e.g., response rate, effort in generating ratings)?

We observed that our prototype messages could achieve comparable performance to regular reminders in terms of response rates. This was slightly surprising because the participants had not come across similar frameworks like ours before, and some of them found our framework confusing at the beginning. Even with these initial setbacks, our proposed messages were able to initiate a comparable number of responses in total. Qualitative data from participants indicated that having an imputed rating made it easier for them to provide a rating because they could use the imputed rating as a reference value. In the absence of an imputed rating in regular reminders, the activity of rating a message was perceived to be more burdensome by people. Prior literature also supports this observation since seeing a reference value can lower user burden by reducing the cognitive load of having to generate their own ratings (Shah and Oppenheimer 2008).

Although we hoped to provide more autonomy to users by involving them in the imputation process (Ryan and Deci 2000; Cai, Jongejan, and Holbrook 2019), some participants from the text message deployment felt that they had less control over the interaction. Such feeling may be induced in cases when people fail to see the benefits of the framework, specifically when they fail to understand why the message was showing an imputed rating and why they should give ratings (i.e., increase users' competence; Ryan and Deci 2000). The explanations designed by us provided some understanding in this context, however, future works on transparency could improve their quality by providing concrete examples (Tintarev 2007). One relevant suggestion that came from our participants is to concretely specify the source of the imputed rating (e.g., 'this is what students felt last year.').

Our investigation of absolute differences between actual and imputed ratings also provided further insights into people's response behavior. We observed that people provided more responses than expected when imputed ratings were close to their actual ratings; but as the differences tended to increase, people's response rates tended to drop. This observation might indicate that when a user does not respond to an assumption message, there is a high probability that the assumed ratings are significantly different from what the user feels. However, looking at the high number of user ratings as similar to imputed ratings, one may also suggest that communicating imputations might introduce a bias. We discuss this potential issue in the next section.

Potential Challenges in Deploying the Framework

RQ2: What potential challenges might arise when users interact with the framework and correct the imputed value?

The findings from our deployment indicate that providing assumptions may bias users toward the assumption. For instance, we observed that participants provided the same rating as imputed value more frequently than what was expected. In some ways, this is a surprising finding since one might expect that users would be more likely to correct assumptions that were very wrong (i.e., different from their actual ratings), and not bother to respond when assumptions were correct. Instead, it seems likely that the accuracy of the assumption was not a large factor in determining participants' response rates (after all, we did ask participants to respond regardless of accuracy). The bias we do see is consistent with the anchoring effect, which states that people's judgments are biased toward the initially presented reference point (Bahník, Englich, and Strack 2016; Furnham and Boo 2011; Grgić-Hlača, Castelluccia, and Gummadi 2022). In other words, the assumption we sent to participants is the reference point, and it anchors participants' ratings so that these ratings converge towards the assumed rating. Literature has proposed different explanations for the anchoring effect, including insufficient adjustments (Kahneman et al. 1982) and retrieval of anchor-relevant information (Chapman and Johnson 1999). Both of them suggest that the anchoring bias could be a result of a lack of thoughtful and effortful information processing (Wegener et al. 2010; Draws et al. 2021). This is an important consideration when considering ways to communicate imputation strategies. However, we do find that the net bias across all users introduced this way was near 0, suggesting that for aggregate results (e.g., mean response value) this impact may not be as important.

Implications

Our results have important implications for how to collect and impute data, especially when using interactive means such as text messages. First, we find that engaging users in reasoning about and correcting imputed data did *not* discourage them from responding (compared to a reminder message). This suggests that researchers and practitioners who value high response rates can still meaningfully engage respondents with the imputation process, as we do in this paper, to enhance the transparency and scrutability of their process. Second, our results reveal a potential danger in doing so – anchoring bias – but our initial data also may be useful in *correcting* for this bias, either by focusing on aggregate measures (e.g., mean), which are less affected by it, or by attempting to account for such bias in the analysis.

Limitations

This work introduces and explores a new framework, which, in virtue of its exploratory nature, has limitations that might be addressed in future work. The absence of statistically significant results in the experiments may be due to a small sample size, and increasing the sample size through a preplanned power analysis may provide a more comprehensive and confirmatory view of the results. Additionally, the generalizability of the findings might be limited as the participants were all residing in North America, and the results may only reflect their perspective. Further studies with diverse populations can provide deeper insights and increase the generalizability of the findings.

Opportunities for Future Work

Based on our findings, below we describe three opportunities for future work.

Exploring potential bias induced by seeing imputations While existing multiple imputation techniques (Seaman, Bartlett, and White 2012; Austin et al. 2021) are well studied in the statistical literature, and their bias under certain circumstances, e.g., *missing-not-at-random* (Resseguier, Giorgi, and Paoletti 2011), is well recognized, potential bias induced by the proposed framework is less known. While in principle, only the agreed and user-adjusted 'imputed' values could be utilized, further work should investigate whether seeing the imputed value may lead to an unconscious deviation from the actual individual rating as a form of cognitive bias (Azzopardi 2021; Draws et al. 2021). Additionally, the nature of induced bias should be explored in the context of different applications (e.g., clinical tools, recommender systems) or data types (continuous versus discrete, high versus low rates of missing data). These explorations will enable system designers to identify potential trade-offs of deploying our proposed framework in various contexts.

Ensuring user autonomy and control Our framework was designed to increase user autonomy and control over the imputation process by allowing them to correct imputations (Ryan and Deci 2000; Alvarez-Melis et al. 2021; Cai, Jongejan, and Holbrook 2019). Nonetheless, during our study, some participants expressed that seeing the imputed ratings made them feel like they lacked control over the interaction. To address this concern, we suggest an alternative approach where the framework would first ask the user if they would like to see the imputed rating. This approach would respect the user's autonomy and only present the imputed ratings if the user consents. However, this would also add an extra step in the process, which could be perceived as overwhelming by some users (Bhattacharjee et al. 2022). These trade-offs highlight opportunities for future research where the optimal level of user autonomy and involvement can be explored and experimented with (Jo et al. 2023).

Informing users of the process and purpose of imputations Throughout the paper, we received diverse suggestions on communicating the process and purpose of imputations. Our experiments in the text message deployment may have provided statistically significant results if we broke down our texts into two or three chunks (Bhattacharjee et al. 2022). There were diverse opinions regarding the choice of wording as well; some people found the word 'assume' appropriate, while others opted for 'guess' or 'think'. Participants also suggested multiple follow-up messages, the first of which could be a regular reminder with no imputation. Future works should experiment with these different options regarding follow-up messages, wording, and sentence structure to personalize the user experience of communicating imputed ratings and allowing users to edit the imputation (Kim et al. 2022). However, careful considerations should be taken to reduce the risks associated with ill-designed and incorrect explanations (Ehsan et al. 2021; Yacoby et al. 2022).

Conclusion

User non-response and the resulting missing data reduce many systems' ability to improve themselves and deliver optimized services. While most of the current approaches to addressing missing data problems do not inform users of the imputation process, we, in contrast, propose a novel framework that communicates the imputed ratings to users and lets them correct the imputations if necessary. Our exploratory work with professionals allowed us to identify several dimensions of the design space, which eventually structured our subsequent deployment through a text messaging probe. Our results revealed that our proposed framework is comparable with regular reminder messages in terms of collecting ratings from users, and seeing an imputed rating eases the burden involved in generating ratings. We also reported that communicating imputed ratings can bias users' responses, indicating the need for future work to improve the framework in several directions. Our work takes a major first step at designing transparent and scrutable frameworks that involve users in the imputation process.

Acknowledgments

We extend our gratitude to Dr. Q. Vera Liao (Microsoft) for her valuable input. Special thanks go to Pan Chen, Karrie Chou, Bingcheng Wang, Linjia Zhou, and Michelle Lin from the University of Toronto for their assistance in data collection. This work was supported by grants from the Office of

Naval Research (N00014-18-1-2755, N00014-21-1-2576), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-06968), and the National Science Foundation (2209819), and a gift from the Microsoft AI for Accessibility program (http://aka.ms/ai4a).

References

Ajzen, I. 2005. *EBOOK: Attitudes, Personality and Behaviour*. McGraw-hill education (UK).

Alvarez-Melis, D.; Kaur, H.; Daumé III, H.; Wallach, H.; and Vaughan, J. W. 2021. From human explanation to model interpretability: A framework based on weight of evidence. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 8, 3.

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–13.

Andridge, R. R.; and Little, R. J. 2010. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1): 40–64.

Austin, P. C.; White, I. R.; Lee, D. S.; and van Buuren, S. 2021. Missing data in clinical research: a tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37(9): 1322–1331.

Azzopardi, L. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*, 27–37.

Bahník, Š.; Englich, B.; and Strack, F. 2016. Anchoring effect. In *Cognitive Illusions*, 223–241. Psychology Press.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, 2–11.

Bhattacharjee, A.; Bayzid, M.; et al. 2020. Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. *BMC genomics*, 21(1): 1–14.

Bhattacharjee, A.; Pang, J.; Liu, A.; Mariakakis, A.; and Williams, J. J. 2023a. Design implications for one-way text messaging services that support psychological wellbeing. *ACM Transactions on Computer-Human Interaction*, 30(3): 1–29.

Bhattacharjee, A.; Williams, J. J.; Chou, K.; Tomlinson, J.; Meyerhoff, J.; Mariakakis, A.; and Kornfield, R. 2022. " I Kind of Bounce off It": Translating Mental Health Principles into Real Life Through Story-Based Text Messages. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–31.

Bhattacharjee, A.; Williams, J. J.; Meyerhoff, J.; Kumar, H.; Mariakakis, A.; and Kornfield, R. 2023b. Investigating the Role of Context in the Delivery of Text Messages for Supporting Psychological Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Cai, C. J.; Jongejan, J.; and Holbrook, J. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, 258–262.

Chapman, G. B.; and Johnson, E. J. 1999. Anchoring, activation, and the construction of values. *Organizational behavior and human decision processes*, 79(2): 115–153.

Chou, Y.-L.; Lin, Y.-H.; Lin, T.-Y.; You, H. Y.; and Chang, Y.-J. 2022. Why Did You/I Read but Not Reply? IM Users' Unresponded-to Read-receipt Practices and Explanations of Them. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–15.

Chounta, I.-A.; and Nolte, A. 2022. The CAT Effect: Exploring the Impact of Casual Affective Triggers on Online Surveys' Response Rates. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–13.

Cooper, H. E.; Camic, P. M.; Long, D. L.; Panter, A.; Rindskopf, D. E.; and Sher, K. J. 2012. *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319–340.

Doan, A. 2018. Human-in-the-loop data analysis: a personal perspective. In *Proceedings of the workshop on human-in-the-loop data analytics*, 1–6.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Du, J.; Hu, M.; and Zhang, W. 2020. Missing data problem in the monitoring system: A review. *IEEE Sensors Journal*, 20(23): 13984–13998.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Feroz, A.; Abrejo, F.; Ali, S. A.; Nuruddin, R.; and Saleem, S. 2019. Using mobile phones to improve young people's sexual and reproductive health in low-and middle-income countries: a systematic review protocol to identify barriers, facilitators and reported interventions. *Systematic reviews*, 8(1): 1–7.

Figueroa, C. A.; Aguilera, A.; Chakraborty, B.; Modiri, A.; Aggarwal, J.; Deliu, N.; Sarkar, U.; Jay Williams, J.; and Lyles, C. R. 2021. Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. *Journal of the American Medical Informatics Association*, 28(6): 1225–1234.

Figueroa, C. A.; Deliu, N.; Chakraborty, B.; Modiri, A.; Xu, J.; Aggarwal, J.; Jay Williams, J.; Lyles, C.; and Aguilera, A.

2022. Daily Motivational Text Messages to Promote Physical Activity in University Students: Results From a Microrandomized Trial. *Annals of Behavioral Medicine*, 56(2): 212–218.

Fogg, B. J. 2009. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, 1–7.

Furnham, A.; and Boo, H. C. 2011. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1): 35–42.

Grgić-Hlača, N.; Castelluccia, C.; and Gummadi, K. P. 2022. Taking Advice from (Dis) Similar Machines: The Impact of Human-Machine Similarity on Machine-Assisted Decision-Making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 74–88.

Groves, R. M.; Presser, S.; and Dipko, S. 2004. The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68(1): 2–31.

Haarhoff, B.; and Thwaites, R. 2015. *Reflection in CBT*. Sage.

Halbesleben, J. R.; and Whitman, M. V. 2013. Evaluating survey quality in health services research: a decision framework for assessing nonresponse bias. *Health services research*, 48(3): 913–930.

Hassenzahl, M. 2008. User experience (UX) towards an experiential perspective on product quality. In *Proceedings of the 20th Conference on l'Interaction Homme-Machine*, 11–15.

Hawthorne, G.; Hawthorne, G.; and Elliott, P. 2005. Imputing cross-sectional missing data: Comparison of common techniques. *Australian and New Zealand Journal of Psychiatry*, 39(7): 583–590.

Jaffe, R.; Nash, R. A.; Ash, R.; Schwartz, N.; Corish, R.; Born, T.; Lazarus, H.; et al. 2006. Healthcare transparency: opportunity or mirage. *Journal of Management Development*.

Jerez, J. M.; Molina, I.; García-Laencina, P. J.; Alba, E.; Ribelles, N.; Martín, M.; and Franco, L. 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2): 105–115.

Jo, E.; Epstein, D. A.; Jung, H.; and Kim, Y.-H. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.

Kahneman, D.; Slovic, S. P.; Slovic, P.; and Tversky, A. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kalton, G.; and Kish, L. 1984. Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, 13(16): 1919–1939.

Kaye, M. 2017. *Stress: The Psychology of Managing Pressure*. DK Publishing. ISBN 9781465464309.

Khan, Z. Y.; Niu, Z.; Sandiwarno, S.; and Prince, R. 2021. Deep learning techniques for rating prediction: a survey of the state-of-the-art. *Artificial Intelligence Review*, 54(1): 95–135.

Khandkar, S. H. 2009. Open coding. *University of Calgary*, 23: 2009.

Kim, T.; Kim, H.; Lee, H. Y.; Goh, H.; Abdigapporov, S.; Jeong, M.; Cho, H.; Han, K.; Noh, Y.; Lee, S.-J.; et al. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. In *CHI Conference on Human Factors in Computing Systems*, 1–20.

Kornfield, R.; Meyerhoff, J.; Studd, H.; Bhattacharjee, A.; Williams, J. J.; Reddy, M.; and Mohr, D. C. 2022. Meeting Users Where They Are: User-centered Design of an Automated Text Messaging Tool to Support the Mental Health of Young Adults. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–16.

Krause, J.; Perer, A.; and Ng, K. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5686–5697.

Law, E.; Yin, M.; Goh, J.; Chen, K.; Terry, M. A.; and Gajos, K. Z. 2016. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4098–4110.

Liao, Q. V.; Zhang, Y.; Luss, R.; Doshi-Velez, F.; and Dhurandhar, A. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 147–159.

MacDonald, S. E.; Newburn-Cook, C. V.; Schopflocher, D.; and Richter, S. 2009. Addressing nonresponse bias in postal surveys. *Public Health Nursing*, 26(1): 95–105.

Mahmoud, D. S. 2021. *Towards Scrutable Decision Tree-based User Model utilising Interactive and Interpretable Machine Learning (SUM-IML)*. Ph.D. thesis, University of Trinity College Dublin.

Montes, G. C.; and Luna, P. H. 2021. Fiscal transparency, legal system and perception of the control on corruption: empirical evidence from panel data. *Empirical economics*, 60(4): 2005–2037.

Niemiec, R. M. 2013a. *Mindfulness and character strengths*. Hogrefe Publishing.

Niemiec, R. M. 2013b. VIA character strengths: Research and practice (The first 10 years). In *Well-being and cultures*, 11–29. Springer.

Pardos, Z. A.; Fan, Z.; and Jiang, W. 2019. Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User modeling and user-adapted interaction*, 29(2): 487–525.

Posner, K.; Brent, D.; Lucas, C.; Gould, M.; Stanley, B.; Brown, G.; Fisher, P.; Zelazny, J.; Burke, A.; Oquendo, M.; et al. 2008. Columbia-suicide severity rating scale (C-SSRS). *New York, NY: Columbia University Medical Center*, 10.

Quiroga, L. M.; Crosby, M. E.; and Iding, M. K. 2004. Reducing cognitive load. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 9–pp. IEEE.

Rader, E.; Cotter, K.; and Cho, J. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–13.

Resseguier, N.; Giorgi, R.; and Paoletti, X. 2011. Sensitivity analysis when data are missing not-at-random. *Epidemiology*, 22(2): 282.

Ricciardelli, E.; and Biswas, D. 2019. Self-improving chatbots based on reinforcement learning. In *4th Multidisciplinary Conference on Reinforcement Learning and Decision Making*.

Richman, M. B.; Trafalis, T. B.; and Adrianto, I. 2009. Missing data imputation through machine learning algorithms. In *Artificial intelligence methods in the environmental sciences*, 153–169. Springer.

Robotham, D.; and Julian, C. 2006. Stress and the higher education student: a critical review of the literature. *Journal of further and higher education*, 30(02): 107–117.

Rogelberg, S. G.; and Stanton, J. M. 2007. Introduction: Understanding and dealing with organizational survey nonresponse. *Organizational research methods*, 10(2): 195–209.

Rothbaum, B. O.; Meadows, E. A.; Resick, P.; and Foy, D. W. 2000. Cognitive-behavioral therapy. In *Effective Treatments for PTSD.*, 320–325. The Guilford Press.

Ryan, R. M.; and Deci, E. L. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1): 68.

Seaman, S. R.; Bartlett, J. W.; and White, I. R. 2012. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, 12(1): 1–13.

Segijn, C. M.; Strycharz, J.; Riegelman, A.; and Hennesy, C. 2021. A Literature Review of Personalization Transparency and Control: Introducing the Transparency–Awareness–Control Framework. *Media and Communication*, 9(4): 120–133.

Shah, A. K.; and Oppenheimer, D. M. 2008. Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, 134(2): 207.

Stout, P. A.; Villegas, J.; and Kim, H. 2001. Enhancing learning through use of interactive tools on health-related websites. *Health Education Research*, 16(6): 721–733.

Suh, H.; Shahriaree, N.; Hekler, E. B.; and Kientz, J. A. 2016. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 3988–3999.

Tétard, F.; and Collan, M. 2009. Lazy user theory: A dynamic model to understand user selection of products and services. In *2009 42nd Hawaii International Conference on System Sciences*, 1–9. IEEE.

Tintarev, N. 2007. Explaining recommendations. In *International Conference on User Modeling*, 470–474. Springer.

Wegener, D. T.; Petty, R. E.; Blankenship, K. L.; and Detweiler-Bedell, B. 2010. Elaboration and numerical anchoring: Breadth, depth, and the role of (non-) thoughtful processes in anchoring theories. *Journal of Consumer Psychology*, 20(1): 28–32.

White, I. R.; and Carlin, J. B. 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29(28): 2920–2931.

Woźniak, P. W.; Kucharski, P. P.; de Graaf, M. M.; and Niess, J. 2020. Exploring Understandable Algorithms to Suggest Fitness Tracker Goals that Foster Commitment. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 1–12.

Yacoby, Y.; Green, B.; Griffin Jr, C. L.; and Doshi-Velez, F. 2022. "If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 219–230.

Yang, Q.; Steinfeld, A.; Rosé, C.; and Zimmerman, J. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13.

Zhou, J.; and Chen, F. 2018. *Human and Machine Learning*. Springer.