Evaluating the Accuracy of User Values Elicited through a Textual Interface

Conducting a user study with a textual interface using questions in isolation to capture user values

Pien Kastelein[1]

Supervisor(s): Catholijn Jonker[1]**, Pei-Yu Chen**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Pien Kastelein
Final project course: CSE3000 Research Project
Thesis committee: Catholijn Jonker, Pei-Yu Chen, Stephanie Wehner

Abstract

This research paper focuses on the accuracy and limitations of user values elicited through a textual interface with questions asked in isolation. The primary objective was to conduct a user study using a textual interface that uses questions in isolation to assess the effectiveness and accuracy of this interface type and questioning style. The study involved exploring various scenarios and associated user values, as well as comparing the textual interface with graphical and audio interfaces tested in four related studies. The user study consisted of 15 participants who interacted with the textual interface. Afterwards, they were tasked with judging and adapting their behaviour models while also evaluating the interface's usability. The findings indicate that while the textual interface demonstrates decent usability, participants did not perceive a strong need for the current system and that compared to other interface types, the textual interface does not yield the most accurate results. This research provides insights into the usability and limitations of a textual interface for eliciting user values. It emphasizes the need for further exploration and development of alternative interface types to enhance accuracy and user engagement.

1 Introduction

Behaviour support applications have gained significant interest and recognition in recent years due to their potential in providing personalized and flexible support to users [10]. To effectively assist users, these applications must understand user preferences, consider contextual factors, and make decisions based on user values. User models that incorporate specific user values have been shown to enhance the performance of behaviour support agents across various domains, including health behaviour change, games, and education [5; 6; 12]. However, a major challenge for these current applications is to accurately and efficiently elicit and update these user values in real-time.

The primary focus of this research paper is to address the question: "Are the user values elicited by a textual interface accurate?". The main objective is to conduct a user study with a textual interface that asks questions in isolation and analyse its results to find out the effectiveness and accuracy of using this type of interface and questioning style. To achieve the research goal, possible scenarios and user values will be identified and defined. This will involve exploring various scenarios and understanding the values associated with them. Secondly, the focus will be on determining the most user-friendly and effective way to present the textual interface. Thirdly, the concept of questions in isolation will be explored as a means to elicit user values. The purpose is to understand how these questions can effectively capture user values. Fourthly, a framework will be developed to measure the effectiveness of the textual interface. This will involve defining appropriate metrics and evaluation methods. Finally, potential limitations related to using a textual interface will be identified and analyzed.

The main contributions of this research include insights into the accuracy of user values elicited through a textual interface and an understanding of the limitations and challenges of using such an interface. This research also explores other aspects of using a textual interface for value elicitation, such as the utilization of questions in isolation and the overall user-friendliness and effectiveness of the interface. These contributions aim to provide a starting point for further research in personalized behaviour support applications.

The remainder of this paper is structured into the following sections: Section 2 presents the background and previous research done regarding value elicitation. This section will also address relevant research, but out of scope for this study. Section 3 describes the methodology and design of the user study, including a description of the textual interface, the evaluation method, and the data collection process. Section 4 will present the results of the user study and Section 5 discusses the findings of the user study along with its limitations. Section 6 will look deeper into the way the user study has been conducted and how this follows the rules for Responsible Research and finally, Section 7 presents the conclusions of this research and recommendations for future research.

2 Background and related work

This background information section of the research paper provides an overview of the previous research conducted in the field, highlights the specific aspects that are relevant to the current study, and identifies important research that is out of scope or not applicable.

2.1 Previous research on value elicitation

The field of value elicitation and behaviour modelling has been the subject of interest for some time, with numerous studies exploring different methodologies and approaches. In this section, we provide an overview of key findings and methodologies from previous research that are relevant to the current study.

Cranefield et al. [5] conducted a study on value-based plan selection in BDI (Belief-Desire-Intention) agents using behaviour tree modelling. Their research focused on using values in the decision-making process of intelligent agents and selecting appropriate plans based on these values. The paper by Berkaa et al. [2] on misalignment in user model elicitation via conversational agents, contributes to the understanding of user behaviour modelling and using behaviour trees. The study also shows the challenges and misalignments that can arise when trying to elicit user values through a conversational interface. In the study conducted by Tielman et al [12], behaviour trees are utilized to derive norms from actions, values, and context. The research helps to understand how these norms can be derived by considering the underlying values and the context in which the actions occur, indicating that the environmental context plays a crucial role in how values are influenced.

In terms of user value elicitation and behaviour modelling using different kinds of interfaces, there were limited pa-

pers regarding the use of textual interfaces. However, studies have explored other types of interfaces, such as graphical interfaces, serious games, the previously mentioned conversational interface and general electronic partners for value elicitation[1; 2; 10]. These studies emphasize the significance of interactive and engaging methods for capturing user behaviour and values.

Unfortunately, limited research exists on the use of questioning in isolation. However, Berkaa et al. [2] found that users initially experienced confusion with certain questions in their study on conversational interfaces. This highlights the importance of formulating clear questions in the textual interface of the current research, to ensure comprehension.

2.2 Related work

While Schwartz's work on values [11] is highly relevant to the field of value elicitation, it is important to acknowledge that it is not directly applicable or included in this current research. Schwartz's theory of basic human values provides a comprehensive framework for understanding and categorizing values based on underlying motivations. However, this study focuses on specific user values within the domain of health behaviour and the utilization of behaviour trees for value elicitation. As Schwartz's work does not provide this level of specificity required for the current research objectives, it is considered out of scope.

3 Methodology

This section will describe the method used for the identification of relevant user values and scenarios, the design and development of the textual interface and the questions in isolation, the evaluation methods of the results to measure the accuracy of the user value elicitation process, the preparations for the user study, and the procedure of the user study.

3.1 Identification of relevant user values and scenarios

The identification of relevant user values and scenarios was a collaborative effort between four other related studies [7; 9; 13; 14]. The aim was to determine the values that are particularly relevant in the context of health and to create scenarios that would indicate potential misalignments between these values and health-related actions.

The main value considered in this research is Health, and it is accompanied by several other values, namely Enjoyment, Comfort, Wealth, Career, Social Acceptance, and Safety. These values were selected based on their potential influence on the choices an individual would make to consider unhealthier choices over healthier choices.

To generate relevant scenarios that would illustrate the misalignment between values and health-related actions, five bigger scenarios were created, each with multiple smaller misalignment scenarios. These scenarios describe various aspects of health improvement, including drinking more water, exercising, eating more healthily, maintaining a better sleep schedule, and improving mental health. After discussion, the four scenarios described in the tables of Table 1 had been chosen to be used for the user experiment. A scenario exists of the following components:

	Scenario 1	Scenario 2
Main Goal	Improve Health	Improve Health
Personal Goal	Increase water intake	Run 3km daily
Alternative	Drink sugary drinks	Watch a movie
Context	Attending a party	Bad weather
Affected Values	Health, Enjoyment, Social Acceptance	Health, Enjoyment, Safety, Comfort

(a) *Summarized scenario descriptions of Scenario 1 and Scenario 2.*

	Scenario 3	Scenario 4
Main Goal	Improve Health	Improve Health
Personal Goal	Maintain more nutritious diet	Improve sleep schedule
Alternative	Eat fast-food	Stay up late
Context	Dining with friends at a restaurant	Work deadlines
Affected Values	Health, Enjoyment, Social Acceptance, Wealth	Health, Wealth, Career

(b) *Summarized scenario descriptions of Scenario 3 and Scenario 4.*

Table 1: *Summarized scenario descriptions*

- **The Main Goal.** This is the same for all scenarios, which is Health improvement.
- **The Personal Goal.** The action the user takes to reach their Main Goal.
- **The Alternative.** The action which is in direct contrast to the Personal goal. It is an action that does not help to reach the Main Goal.
- **The Context.** A situation or reason why the user would choose the Alternative
- **The Affected Values.** All values that are affected either positively or negatively when choosing the Personal Goal or the Alternative.

3.2 The textual interface prototype and questions

To save development time and ensure a user-friendly experience, the online chatbot builder, Landbot [8], was chosen for creating the textual interface. Landbot is a platform that allows users to create interactive conversational interfaces. It can be accessed online, making it convenient for participants to fill out and providing easy data collection and storage for efficient result management. Using a chatbot offers the advantage of simulating a conversation rather than a form-based interaction, creating a more engaging and personalized experience. The clean and minimalist design of the chatbot interface that can be seen in Figure 1 was deliberately chosen to minimize distractions and maintain user focus on answering the questions effectively.

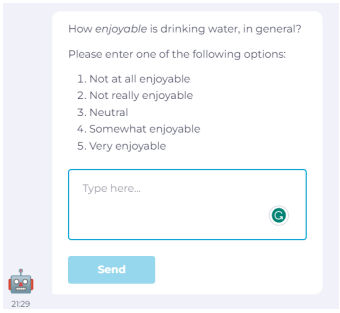


Figure 1: *Textual interface prototype, made using Landbot chat builder.*

The formulation of the questions for each scenario has

been done in collaboration with the related studies [7; 9; 13; 14]. The questions in this textual interface were designed to be asked in isolation, focusing on a single value in a specific situation. As an example, let's consider Scenario 1, where the Personal Goal is to increase water intake. One of the questions asked in the textual interface was: "*How enjoyable is drinking water, in general?*" The participants were instructed to select one of the following options: "*Not at all enjoyable*," "*Not really enjoyable*," "*Neutral*," "*Somewhat enjoyable*," or "*Very enjoyable*." By asking questions in isolation, the participants are forced to think about that specific value in that specific situation. For each general and context-based situation, a question about the Personal Goal and the Alternative must be asked, resulting in a total of four questions per Affected Value.

3.3 Evaluation methods of collected results

The participants' choices and responses were used to develop and compare user models, representing their behaviour within behaviour trees. The general, context-based and user-optimized user values were compared using two different strategies to measure the accuracy of the textual interface in capturing and representing user values.

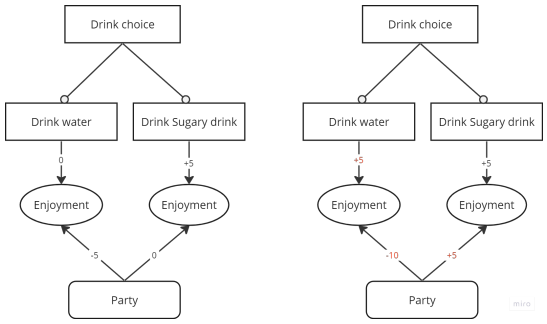


Figure 2: *Simplified user model, modelling only the Enjoyment of drinking water or soda given the context of a party with values changed by the user in the left model.*

The first strategy involved calculating the Hamming distance between the user value trees. The Hamming distance measures the dissimilarity between two objects [3]. In this research, these objects are the trees and by counting the number of different values between them the Hamming distance is calculated. In this analysis, one changed value was considered to have a Hamming distance of 1. Consider the two extremely simplified behaviour trees in Figure 2 for Scenario 1. In these trees, only the Enjoyment value is given. For this tree, the participant decided that the values elicited by the interface were not fully correct and changed them. The enjoyment of drinking water with and without the Context of a party has changed, and the enjoyment of drinking soda at a party has changed. This means that the Hamming distance is 3, as the user changed three values.

The second strategy involved comparing the absolute total changes made between the two trees. For Figure 2, the total change would be $abs(+5) + abs(-5) + abs(+5)$. By combining this strategy with the Hamming distance, it was possible to calculate the average value change per edge within the behaviour tree. For the simplified behaviour tree in Figure 2, the average value change per edge would be $\frac{abs(+5)+abs(-5)+abs(+5)}{3}$.

To further assess the accuracy of eliciting user values, the results of this study were compared with the four related research studies, also part of the TU Delft Research Project [7; 9; 13; 14]. These related studies also explored the accuracy of user value elicitation, only using different questioning contexts or interfaces. To investigate the potential impact of different interfaces on user values, each researcher of the related study participated in all other studies. This way, a comparison of accuracy can be made between these interfaces.

Lastly, to evaluate the usability of the textual interface, the participants were asked to complete a System Usability Scale (SUS) survey in Microsoft Forms. The SUS survey is a reliable usability scale that is often used to assess the usability and user-friendliness of the interface [4]. It consists of a set of statements related to usability and the participants were asked to rate their level of agreement with each statement using a scale from one to five.

3.4 The user study

The user study was conducted in three steps. The first step involved selecting the appropriate participants for the study. The second step focused on preparing and informing the selected participants about the experiment, ensuring they had a clear understanding of the procedures and expectations. Finally, the third step was the actual execution of the experiment.

User study participants

To evaluate the effectiveness of the textual interface, a user study was conducted involving a group of 15 technologically literate individuals between the age of 20 and 65. Out of these participants, 11 exclusively tested the textual interface, while four participants also tested one other textual, two graphical and one audio interface.

User study preparations

Before the start of the experiment itself, each participant was briefed on how long the experiment would take. They were also requested to complete a Human Research Ethics (HREC) form, specially designed for this research and approved by TU Delft. This form ensured that participants provided their informed consent to participate in the study, demonstrating their understanding of the study's purpose, procedures, and potential risks or benefits.

User study procedure

During the study, participants interacted with the interface and were instructed to provide truthful answers to the presented questions. Throughout this interaction, a researcher was present to collect the answers and develop corresponding behaviour trees for each participant, scenario, and context. The actions of the researcher remained out of sight of the participant to ensure unbiased responses. If the participant chose to do the experiment by themselves at home, the researcher remained available via online communication.

Upon answering all the questions within each scenario, it was explained to the participants how the modelling of behaviour trees works. They were presented with their own user models based on their responses in the textual interface. In the cases where participants disagreed with their model, the participant could indicate how and why they would make adjustments according to their perception of their values. The user then changed the values in the initial model and both models were saved.

In the final step of the user study, the participants were requested to complete the SUS survey. The participants who tested all five interfaces were asked to fill out the SUS survey for each interface individually.

4 Results

The first section will present the results of the user study conducted for the textual interface in different formats. First, the individual results per user are presented, then the processed data is presented as statistical data of the collected results.

In the second section, the results of the collaborative study will be presented next to the results of this study. The results of the four participants that participated in all studies are also presented in this section.

The last section will cover the results of the SUS survey.

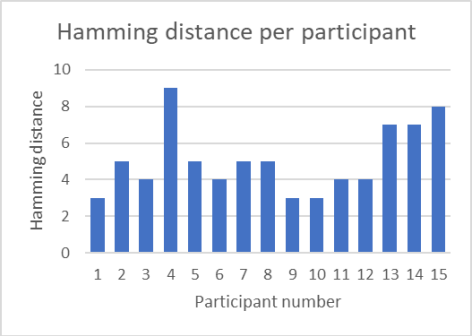
4.1 Results of the user study using the textual interface

Figure 3 displays the Hamming distances, value differences and average value differences per value change of the participants. It is important to note that the participant data is presented in a randomized order.

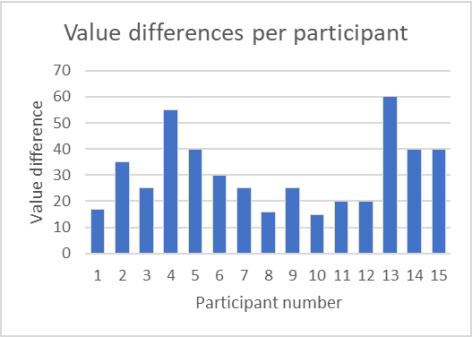
The Hamming distance of participant number 4 is the highest compared to other participants, but in Figure 3c it can be seen that the average value difference per value change for participant number 4 is not significantly higher than that of the other participants. This means they changed relatively more values but with small steps. Interestingly, participant 13 has a very high value difference and also a high value difference per value change. This suggests that this participant made larger changes to their values.

Additionally, participant 8 is noteworthy. This participant has a Hamming distance of 3, a value difference of 16, and an average difference per value change of 3. This participant is one of the participants that did not agree with only changing the values by multiples of 5. They wanted to change their values with multiples of 2, which explains the inconsistency in the data in Figure 3.

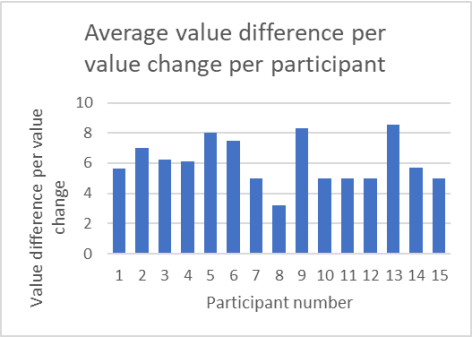
Lastly, some of the calculated statistical data is shown in Table 2. Here can be seen that the mean and median of the Hamming distance and the total average value change per changed value are quite close to each other and that this is not the case for the total value differences. Additionally, Table 2b shows a relatively high standard deviation for the value difference. This can also be seen in Figure 3b, as the results of each participant differ quite a lot compared to Figure 3a and 3c.



(a) *Hamming distances of all participants.*



(b) *Value differences of all participants.*



(c) *Average value difference per changed value of all participants.*

Figure 3: *Collected data of all participants*

Sample Size	Mean	Median	Standard Deviation
15	5.07	5	1.81

(a) *Total average Hamming distance of all results.*

Sample Size	Mean	Median	Standard Deviation
15	30.87	25	13.43

(b) *Total average value difference of all results.*

Sample Size	Mean	Median	Standard Deviation
15	6.09	5.71	1.47

(c) *Total average value difference per changed value of all results.*

Table 2: *Calculated statistical data of all results.*

4.2 Results of the collaborative study

This research was conducted in collaboration with four related studies that focused on evaluating the accuracy of user values elicited through different interfaces and questioning structures. Specifically, this research project focused on a Textual interface with questions in Isolation (TI), while the other studies explored a Textual interface with questions in Comparison (TC), a Graphical interface with questions in Isolation (GI), a Graphical interface with questions in Comparison (GC), and an Audio interface with questions in Isolation (AI). In this section, we present the combined results of the four participants who took part in all five studies, followed by a statistical summary of the individual results for each of the five studies.

Results of the participants participating in all five studies

Firstly, it is important to note that the participants who participated in all studies, were the researchers themselves. Table 3 presents the average results of all interfaces.

The AI yields the highest accuracy, as it shows minimal differences between the original and changed values. On the other hand, the GC shows the most changes, although the average value change per edge is less than the TC.

	Hamming Distance	Value Difference	Value Difference per Changed Value
TI	6.5	40	6.15
TC	6.75	55	8.15
GI	3.5	17.5	5
GC	8	50.5	6.31
AI	0.5	2.5	5

Table 3: *Averages of the data of the four participants who had to do all five interfaces: **T**extual in **I**solation, **T**extual in **C**omparison, **G**raphical in **I**solation, **G**raphical in **C**omparison and **A**udio in **I**solation.*

Individual results of the collaborative studies

The researchers from each study collaborated and shared their respective statistical data. Each study involved 15 participants, with 4 participants overlapping across all studies, and 11 different participants for each individual study. However, not all researchers calculated the average value difference per changed value. As a result, the value differences per changed value in Table 4 are calculated by dividing the mean value differences by the mean Hamming distances, as described in Section 3.3.

	Hamming Distance	Value Difference	Value Difference per Changed Value
TI	5.07	30.87	6.09
TC	0.80	9.67	12.09
GI	1.30	8.00	6.15
GC	5.33	36.87	6.92
AI	3.60	13.50	3.75

Table 4: *Averages of all data of all five Research Projects: **T**extual in **I**solation, **T**extual in **C**omparison, **G**raphical in **I**solation, **G**raphical in **C**omparison and **A**udio in **I**solation.*

From Table 4, it is evident that the TC shows a significantly lower Hamming distance compared to its value in Table 3, making it the interface with the lowest Hamming distance overall. Additionally, the average value difference for TC is considerably lower. However, the value difference per changed value remains the highest among all the interfaces, suggesting that the changes made by users have been significant.

Both textual interfaces perform relatively well, while the GC consistently ranks among the lowest in terms of accuracy across all measurements. On the other hand, the AI shows a decrease in performance compared to its results in Table 3, yet it still maintains the lowest value difference per changed value among all the interfaces.

For a more detailed explanation and the precise data for each individual interface, please refer to the respective papers [7; 9; 13; 14].

4.3 Results of the System Usability Scale survey

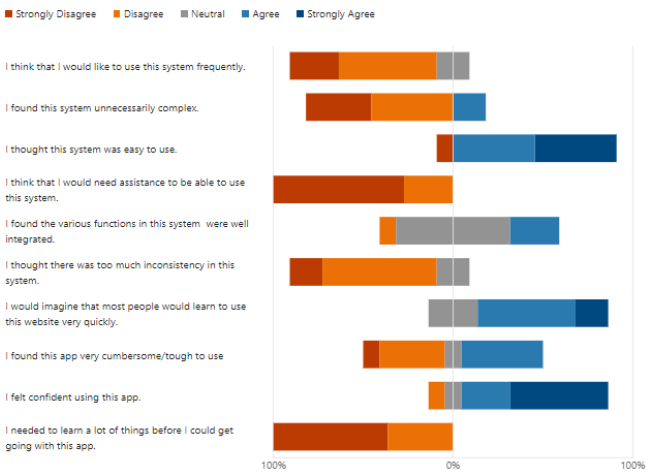


Figure 4: *Results of the System Usability Scale survey.*

The SUS survey was administered using Microsoft Forms, enabling automatic processing and a visual representation of the survey responses as shown in Figure 4. For these results, it is important to note that these are the results of the 11 participants that only used the Textual interface described in these results.

Lowest score	Highest score	Average score
50	87.5	68.9

Table 5: *System Usability Survey (SUS) scores.*

To start with the frequency of system usage, it is clear to see that most users would not use this interface frequently. Further analysis of the other answers reveals that this is not attributed to the complexity of the system. In fact, the majority of users found the interface easy to use, easy to learn, and did not feel the need for assistance. The results in Table 5 show that the SUS score of the interface is slightly above average with an average score of 68.9. The lowest SUS score is 50, indicating that this user did not think the interface was usable to them. The main contributor to the lower scores is that the participants did not perceive a need for this interface in its current form, as can be seen from the first result in Figure 4.

5 Discussion and limitations

This section will provide several key points that emerged from the results of this study and will shed light on the potential implications and limitations of this study.

5.1 Analysis of the results of the user study

Firstly, every user made changes to their behaviour tree when presented with the final model. These changes had an average value difference of approximately 5, which can be seen in Table 2. This suggests that the participants tended to make slight adjustments to align them more closely with their preferences. However, it is possible that the participants did not fully agree with how their behaviour was initially modelled and felt the need to make adjustments based on their personal preferences or understanding. However, it is also a possibility that the way the behaviour tree was explained to the participants during the study might have influenced their perception and prompted them to make modifications.

It is important to note that the data of the value differences observed in Figure 3b and Table 2b has two high outliers. Participant 4 en Participant 13 have a very high value difference compared to the other participants. In Figure 3a it can also be seen that Participant 4 also has the highest Hamming distance. The issue could be that, at the beginning of the research, there were three instances where some values were not stored correctly. Because of this, these values had to be asked while the participant was reviewing their behaviour tree. However, these changes were recorded as "changes" to the original tree, which might have introduced some skew in the data.

Additionally, it is important to note that some participants wanted to change their values in even smaller steps, instead of the predefined increments of 5. This can be seen in Participant 8 in Figure 3. The participant indicated that they wanted to change their values with steps of 2. Allowing for a wider range of value adjustments has the potential to improve the accuracy and personalization of the elicited values. However, it is important to consider that offering a wider range may also introduce challenges in terms of usability and user experience.

5.2 Analysis of the results of the collaborative study

Regarding the comparisons made between the results of the related studies, several observations can be made. First, it is important to consider the difference between the results of the participants who took part in all the studies and the overall results of the four related studies. As shown in Table 3 and Table 4, the audio interface appears to perform the best for the participants who used all the interfaces, while the differences are less visible in the overall results. One potential explanation for this difference is the fact that the participants who used all the interfaces were the researchers themselves. This introduces a potential bias, as these participants already had preconceived notions about how the program should or would work. This could have influenced their interactions with the interfaces and, consequently, skewed the data in some way.

To address the question of determining the "best" interface for eliciting user values, a ranking based on the results presented in Table 4 can provide valuable insights. As discussed in Section 4.2, it seems that the Graphical interface with questions in Comparison performs worst overall, with the highest Hamming distance and value difference. From Table 4 it can also be seen that the Textual interface with questions in Isolation performs well in terms of value difference per changed value. However, this interface ranks relatively lower in other accuracy measures. Ranking the three resulting interfaces becomes more challenging as it depends on which accuracy measure is given the highest priority. In general, the Graphical interface with questions in Isolation appears to yield the best overall results. However, it is important to note that if a specific accuracy measure is considered to be the most important, the ranking of the interfaces could change.

Lastly, based on the data presented in Table 4, certain observations can be made regarding the different interface types and questioning styles. It can be seen that a textual interface tends to benefit from questions in comparison, while the graphical interface shows better performance with questions in isolation. When determining the overall best-performing interface type, it is important to consider that the audio interface has only been tested with questions in isolation, limiting the conclusions that can be drawn from the performance of all interface types. However, from the data given in Table 4, the interfaces that use questioning in isolation perform relatively well. This could be the result of the fact that the interfaces with questions in isolation had more questions to ask than the interfaces with questions in comparison. This could have resulted in a more fine-tuned result, as each value was asked in each situation, instead of the comparison between two situations.

5.3 Limitations

The main limitation of this research was the constrained time frame. With only 8 weeks available, the scope of the study had to be narrower than initially desired. Additionally, the sample size was limited to only 15 participants due to the time, which may not fully represent the diversity of user perspectives. This time limitation also influenced the textual interface, which would have benefited from multiple iterations to refine its design and functionality.

Regarding the usage of the textual interface, there was an issue related to user input. Although the participants were instructed to enter numbers from 1 to 5, they had the freedom to input any value. This resulted in one participant accidentally typing a non-numeric character and being unable to change it. This value was asked again when reviewing the behaviour tree, but this value was also recorded as "change". Providing clearer instructions and implementing input validation mechanisms could help prevent such issues and improve the user experience.

Some participants expressed difficulties in understanding certain questions and how their responses would impact the values in a given situation. This led to confusion and, in some cases, participants resorted to selecting the "*Neutral*" option due to the lack of comprehension. To improve the effectiveness of the textual interface, careful attention should be paid to question clarity and ensuring that users fully grasp the context and implications of their responses.

During the comparison with the other related studies, the

issue arose that the questions asked in each interface differed more than was intended. While the scenarios remained consistent across different interfaces, the differences in question formulation may limit the overall comparability of the results. Additionally, although the studies followed a similar structure, there were still variations that influenced participants' perceptions and potential adjustments to their behaviour models. To enhance the validity and reliability of future research comparisons, it would be beneficial to establish guidelines for question formulation and the overall user study methodology. This would contribute to a more consistent and meaningful comparison of results across different studies.

A notable observation across all interfaces was participant fatigue. Some participants expressed boredom or perceived the questions as time-consuming, resulting in less motivation and attention towards the end of the study. Addressing participant fatigue and maintaining engagement throughout the interaction are important considerations for future research to ensure high-quality data collection.

Lastly, it is essential to acknowledge again the potential bias introduced by the fact that the participants who used all interfaces were the researchers themselves. This bias should be taken into account when interpreting the results.

6 Responsible Research

In conducting this research, several measures were taken to ensure the integrity and ethical conduct of the study. Firstly, all data collected during the research process is accessible and was used exclusively for the purpose of conducting this study. Participants' responses and data were anonymized and treated with strict confidentiality.

To promote reproducibility, the research methodology and analysis procedures were thoroughly described. Detailed documentation of the scenarios and procedures can be provided to enable other researchers to replicate the study's findings and facilitate further research in the field. This transparency allows for the validation and verification of the study's results.

In terms of ethical considerations, participants were provided with an HREC form specifically designed by TU Delft for this research. They were required to read and sign this form, indicating their informed consent to participate in the study. The HREC form outlined the purpose, procedures, and potential risks and benefits of the research, allowing participants to make an informed decision about their involvement.

Lastly, all findings and results have been included in this research. No data has been held back and all data can be provided for further research. By adopting all of these practices, this study aims to be transparent, reproducible and valid.

7 Conclusions and Future Work

This research aimed to assess the effectiveness of using a textual interface for eliciting human values through a user study. The accuracy of eliciting human values is crucial, especially in the context of health applications that strive to provide personalized experiences. Understanding individuals' values and how they may vary in different situations or scenarios is essential for tailoring these applications to meet users' specific needs.

The research began by creating four distinct scenarios that aimed to identify values relevant to the context of health. These scenarios presented situations where individuals would ideally choose the healthier option but might opt for the less healthy alternative due to specific contextual factors. These scenarios served as the basis for value elicitation in the next stages of the research.

In the next step, a textual interface was developed with a user-friendly and minimalist design. The questions in the interface were designed to be in isolation. The questions were phrased to focus on one aspect of the scenario at the time. This approach aimed to provide clear and focused responses from participants, enabling accurate value elicitation.

The textual interface was then tested with a group of 15 participants, ranging in age from 20 to 60 years. Each participant interacted with the interface individually and evaluated its user-friendliness. The collected data from these sessions were processed to create user models for each participant, capturing their values based on their responses. To determine the accuracy of the user models, participants were asked to review and change their own models if they felt they needed to.

The results revealed that all participants wanted to change certain aspects of their user models. This suggests that while the interface succeeded in eliciting general positive or negative values, it was not able to capture the exact user values, according to the participants.

Furthermore, a significant majority of participants indicated that they did not perceive a current need for the program in its existing state. This emphasizes the need for further research and development to enhance the effectiveness of textual interfaces for value elicitation.

In addition to the conducted user study, a comparison was made with four related studies that employed different interface types and questioning methods. The comparison revealed that a graphical interface with questions in isolation tended to elicit user values with the highest accuracy based on the three different accuracy measurements used. However, both an audio interface with questions in isolation and a different textual interface also demonstrated favourable results in terms of accuracy. It is important to note that the current findings do not provide a definitive conclusion on the most accurate interface and questioning approach on a larger scale, as in all studies the participant groups only existed of 15 participants. This limitation indicates a need for further research with more participants to be able to draw more conclusive findings.

Lastly, it is important to address the limitations identified in this study. These include the constraints on the total available time, instances of faulty user input, challenges with question clarity, a fault in data collection, participant fatigue, and the need for standardizing question formulation across studies. To overcome these limitations, future research should consider allocating more time, improving user input mechanisms, enhancing question clarity, implementing robust data collection procedures, addressing participant fatigue, and including a larger and more diverse participant pool. These measures

will contribute to the reliability and broader applicability of the study's findings.

In conclusion, this research contributes valuable insights into the accuracy and limitations of a textual interface for value elicitation. It serves as a stepping stone for future researchers to refine and expand upon these findings. By addressing the identified limitations and building upon the knowledge gained, other researchers can continue enhancing the effectiveness of personalized support systems.

References

[1] J. Almeida. *Serious Games as a Behaviour Elicitation Tool: Applications to Evacuation Scenarios.* PhD thesis, 01 2016.

[2] J. Berka, J. Balata, C. Jonker, Z. Mikovec, M. Riemsdijk, and M. Tielman. Misalignment in semantic user model elicitation via conversational agents: A case study in navigation support for visually impaired people. *International Journal of Human-Computer Interaction*, 38:1–17, 04 2022.

[3] A. Bookstein, V. Kulyukin, and T. Raita. Generalized hamming distance. *Information Retrieval*, 5, 10 2002.

[4] J. Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[5] S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum. No pizza for you: Value-based plan selection in bdi agents. pages 178–184, 08 2017.

[6] M Kließ, M Stoelinga, and M. B. Riemsdijk. *From Good Intentions to Behaviour Change: Probabilistic Feature Diagrams for Behaviour Support Agents*, pages 354–369. 10 2019.

[7] M. Krupskis. Designing graphical user interface to elicit personal values. 6 2023.

[8] Landbot, https://app.landbot.io/.

[9] S. Mendez. Eliciting personal values through isolation questioning: A graphical interface approach. 6 2023.

[10] M.B. Riemsdijk, C. M. Jonker, and V. Lesser. Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. 2:1201–1206, 01 2015.

[11] S. H. Schwartz. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1), 2012.

[12] M. Tielman, C. M. Jonker, and M. B. Riemsdijk. What should i do? deriving norms from actions, values and context. pages 35–40, 07 2018.

[13] B. Vizuroiu. Accuracy of textual interfaces using comparative questions to elicit personal value-related information. 6 2023.

[14] E. Voorneveld. The accuracy of an audio interface designed for value elicitation. 6 2023.