

**Designing and Evaluating an Inclusive Version of the Bot Usability Scale (BUS) for Interaction with
AI-Based Conversational Agents**

Maria N. Hristova

2274914

Department of Behavioural Management and Social sciences, University of Twente

56604: Psychology

Dr. Simone Borsci

J. Landwehr

05 July 2022

Abstract

Inclusivity and accessibility are words that are heard increasingly often in discussions, especially ones surrounding product design and planning. And with the increasing presence of chatbots and other similar automated customer service assistants, it is important to consider the accessibility factor early on to ensure that future technologies can be used by everyone. People with disabilities are a marginalized group that requires additional care for a proper customer service experience. Thus, the perspective of people from this group is extremely important.

For the world to become more accessible first the issue of accessibility in research needs to be addressed. This suggests the need to provide people with disabilities with an opportunity to participate in research, for which the accessibility of the scales used needs to be considered. The Bot Usability Scale (BUS) that is proposed by Borsci et al., 2022 has the potential to be used as a tool for assessment of the user's experience with various chatbots and has been adapted for multiple languages.

A new design is proposed to replace the existing one for the BUS-11. This new design, BUS-A is based on principles for accessible design and was constructed after testing possible variations with a focus group with representatives from the end-users. The new design makes use of the original 11 item formulation of the BUS since there were no issues found in the phrasing of the items. This proposed design has also been tested with a larger sample of participants and compared to results from the BUS-11 scale to check if the two versions measure the same concept. It was found that the assumed underlying factorial structure of the BUS-11 does not fit the BUS-A, based on unsatisfactory CFI and RMSEA values, therefore it cannot be confidently stated they measure the same concept. However, on its own, the BUS-A showed sufficient internal consistency and reliability comparable to the original design, which suggests the scale has good psychometric properties and can be considered as a potential measurement tool of user satisfaction.

Keywords: Chatbots, Accessibility, Inclusivity, Bot Usability Scale-11, Accessible design

Contents

INCLUSIVE DESIGN IN RESEARCH	4
MEASURING USER EXPERIENCE AND SATISFACTION WITH CHATBOTS	4
GOAL OF THE PRESENT STUDY	6
PHASE 1: DESIGN AND EVALUATION OF AN ACCESSIBLE VERSION OF THE BUS 11: BUS-A	7
DESIGNING AN ACCESSIBLE VERSION OF THE BUS-11	7
METHOD	13
STUDY DESIGN	13
PARTICIPANTS	14
MATERIALS	14
PROCEDURE	15
DATA ANALYSIS	16
RESULTS	16
IMPLICATIONS OF FOCUS GROUP	17
METHOD	20
STUDY DESIGN	20
PARTICIPANTS	21
MATERIALS	21
PROCEDURE	22
DATA ANALYSIS	23
RESULTS	24
DIFFERENCES IN MEAN SCORES BETWEEN CONDITIONS	28
CONFIRMATORY FACTOR ANALYSIS ON A FIVE-FACTOR STRUCTURE	30
DISCUSSION	32
LIMITATIONS	34
CONCLUSION	35
APPENDIX A	40
APPENDIX B	43
APPENDIX C	47
APPENDIX D	48
APPENDIX E	55
APPENDIX F	57
APPENDIX G	63
APPENDIX H	84

Designing and Evaluating an Inclusive Version of the Bot Usability Scale (BUS) for Interaction with AI-Based Conversational Agents

After spending some time on a webpage, it is common for a user to encounter a pop-up window somewhere on the screen asking if they need assistance with a task. Customer services, assistance with shopping, and even educational support have been entrusted to artificially created conversational agents, also known as chatbots, by more than a few companies. A ‘chatbot’ is a rules-based, bounded system that has well-defined actions it can perform (Amelia, 2020). In 2019, over 40 million active businesses across the globe were exchanging over 20 billion messages each month with customers via chatbots (Acquire, 2022) with the top five countries making use of such technology being the USA, India, Germany, the UK, and Brazil. Those numbers are expected to grow and chatbots are regarded as essential for the future. By the end of 2022 chatbots are projected to handle up to 90% of healthcare and banking customer services (Gilchrist, 2017). Yet many technologies still fail to provide baseline accessibility to their users and accommodate people with disabilities (Brewer, 2018).

Chatbots have the possibility to adapt and change even after they have been launched for the public to use. Depending on the purpose of the interaction, they can provide different prompts to users or have new patterns of interaction developed. Behind a chatbot, there is an intelligent algorithm that can conduct text-based conversations with users as a result of training. However, while chatbots do interact with users, there is a distinction to be made between them and Conversational Artificial Intelligence (A.I.). Conversational A.I. systems can account for the human variance in dialogue and therefore are better at handling interactions with more human-like agility. This is hard to achieve with a traditional chatbot that has predetermined response options (Google Dev., 2020). What chatbots can do, however, is to reach out proactively to users and give them solutions to specific issues based on metadata from their environment (Winkler & Soellner, 2018). This prevents spending additional time in search for the right information both for the user and for any potential human support. Furthermore, it makes the user feel personally addressed since chatbots use data specific to the context of the user to provide responses and initiate communication. And with the trends pointing towards wider adoption of the chatbot functionality, it is important to evaluate their usability and the user satisfaction to further improve the experience of all users.

Inclusive design in research

With the expected wider implementation of web-based services such as customer support chatbots or other autonomous conversational agents for various purposes, it is important to consider inclusivity during the design process. As mentioned, chatbots are seeing rapid growth in countries all over the world and they are expected to become even more prevalent in customer services (Acquire, 2022). This suggests even more users engaging with the conversational agents and with this increase we can also expect an increase in the variety of the user's characteristics.

According to the World Health Organization (WHO), about 15% of the world's population live with some form of disability, with 2 – 4 % experiencing significant difficulties in functioning (Carroll, 2012). The official definition of disability provided by WHO includes three aspects: impairment, activity isolation and participation restriction (Centre for Disease Control and Prevention, 2017). Disability can have many forms and it can be seen as the result of people with health conditions experiencing negative interactions with personal or environmental factors (DSM-V, 2013, p. 66). The discussion around inclusivity of people with various levels of disability has been ongoing for a long time. In the United Nations Convention on the Rights of Persons with Disabilities from 2011, society was urged to work alongside persons with disabilities, as their participation is essential to achieving an inclusive worldwide community (*United Nations Convention on the Rights of Persons with Disabilities - Employment, Social Affairs & Inclusion - European Commission*, n.d.). The European Union's equivalent of this convention is the European Accessibility Act (*European Accessibility Act - Employment, Social Affairs & Inclusion - European Commission*, n.d.) which aims to improve the functioning of the products and services internal to the member states. This would further accommodate people with disabilities. The motto "Nothing About Us Without Us" has been used by Disabled People Organizations worldwide and stresses the importance of involving people with disabilities and their feedback (*International Day of Disabled Persons 2004 - United Nations Enable*, n.d.).

Measuring User Experience and Satisfaction with Chatbots

To ensure that future technologies are created with people with disabilities in mind and do not exclude anyone from receiving the best possible service, involvement of people with various limitations is needed during all stages of the design process. One step to making this a reality is by making research more accessible to draw in more participation. To make involvement in research effortless for people with disabilities, the accessibility of questionnaires

needs to be evaluated and improved (Goegan et al., 2018a). One scale that has been designed specifically for use with chatbots is the recently validated Bot Usability scale (BUS-11) (Borsci et al., 2022). Usability, as defined by the ISO 9241-11, is the extent to which a system or service can be used by specified users to achieve given goals with effectiveness, efficiency and satisfaction in the context of use (ISO/IEC JTAG, 2014). Currently, the predominant way of measuring such variables is by either adjusting already existing scales to the context of chatbots or by different measures for satisfaction (Borsci, 2021). The BUS-11 reliably enables end-users to express their perceived experience with a chatbot which is confirmed by a Cronbach's Alpha value for the survey of $\alpha = .90$. As it can be seen in Table 1, the BUS scale consists of 11 items divided among 5 factors. The scale makes use of a standard 5-point Likert scale for answering options from 'Strongly Disagree' to 'Strongly Agree'.

Table 1

Items of the BUS-11 and its five factors

Factor	Item
1 - Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable.
	2. It was easy to find the chatbot
2 - Perceived quality of chatbot	3. Communicating with the chatbot was clear.
	4. The chatbot was able to keep track of context.
	5. The chatbot's responses were easy to understand.
3 - Perceived quality of conversation and information provided	6. I find that the chatbot understands what I want and helps me achieve my goal.
	7. The chatbot gives me the appropriate amount of information
	8. The chatbot only gives me the information I need.
	9. I feel like the chatbot's responses were accurate
4 - Perceived privacy and security	10. I believe the chatbot informs me of any privacy issues

5 - Time response

11. My waiting time for a response from the chatbot was short.

The scale strongly correlates with another measurement of user satisfaction – the Usability Metric for User Experience (UMUX) - Lite. The UMUX-Lite is a standardized measurement related to the perceived ease-of-use and perceived usefulness of a product (Lewis & Sauro, 2020). It is a short two-item scale that can be adapted to various products. The BUS-11, however, unlike other scales of satisfaction, considers aspects such as the accessibility of the chatbot, the time to response and perceived privacy. The scale is available in multiple languages such as English, Dutch, Spanish and German, and it is under validation in Italian and Russian (e.g, Lopez, 2021).

Goal of the Present Study

The present study is divided into two consecutive parts with the overall goal to create an inclusive and accessible version of the BUS-11. In the first stage, the current design of the BUS will be used as a basis, upon which a more accessible version, named BUS-A, will be built. The design will be adjusted and tested with a focus group to obtain meaningful insights to proceed with creating the accessible version design. In the second stage, the BUS-A design will be evaluated for validity and reliability as a scale and whether it can replace the current version will be considered. For this to happen the two designs must prove to measure the same concept of user satisfaction and have the same underlying structure. As the focus of stage one is to implement accessibility in the design and validate it through feedback from potential end-users, two research questions were formulated concerning this:

RQ1: What are design elements that can be implemented to improve the accessibility of a scale design based on scientific literature and design recommendations?

RQ2: Does the current item formulation of the BUS-11 comply with accessibility requirements?

The second phase of the study takes place after the definitive version of BUS-A has been designed and solidified and has the purpose to assess the scale with a larger sample of people. Based on the recommendation from the first phase of this study, the proposed design will be tested with a larger sample to explore if it is suitable for use. In this phase comparison of the

BUS-A against the BUS-11 will be done to establish whether the two scales perform similarly on aspects such as internal reliability and validity, item correlation and factorial structure. Establishing that the novel design does not influence the scale's validity and reliability is the first step toward providing more accessible options to future users. Moreover, if the redesign is validated with more people with disabilities this could contribute to better understanding and possible further improvements of the BUS-A.

To prove that the accessible redesigned version can be confidently used in the future, it must be shown that there are no significant differences between data collected with the new proposed design and with the one that has been used thus far. Therefore, the research question this phase strives to answer is:

RQ3: Is the factorial structure of the BUS preserved in the BUS-A, based on CFI value of 0.90 or greater and RMSEA value of less than 0.08, and is the new design a reliable scale to measure user satisfaction based on Cronbach's Alpha value equal to 0.7 or greater?

Phase 1: Design and evaluation of an accessible version of the BUS 11: BUS-A

Designing an Accessible version of the BUS-11

One general guideline for accessible and inclusive design is provided by W.C.A.G. 2.0 (*W.C.A.G. 2 Documents - Web Accessibility Initiative (WAI) - W3C*, n.d.). There are three distinct levels of conformance with those standards that serve as an indication as to what extent the given content is accessible. The three levels are as follows: A (lowest), AA (mid-range), and AAA (highest) (Global Health Workforce Alliance, 2019; *WCAG 2.0 Conformance Levels - UCOP*, n.d.). Each level in this scale indicates that the requirements for the previous levels have already been met, so for example content marked as meeting conformance level AA in full also completely satisfies all the criteria required for level A. The WCAG standards are developed with the purpose of making content accessible to a wider range of people with varying disabilities. Following these standards is often a way to improve the usability of the content that is being provided (*W.C.A.G. 2 Documents - Web Accessibility Initiative (W.A.I.) - W3C*, n.d.). Since the standards are not technology-specific and can be applied to multiple contexts, this provides the opportunity to use them as a reference when designing even questionnaires. Abiding the WCAG standards (Ribera et al., 2009), the implementation of the new elements was divided into categories that can be seen in Table 2.

This part of the study was done in collaboration with Anna Boyko, another researcher from the University of Twente who also has an interest in improving access to research for everyone. To use the advantage of different perspectives, the two researchers on the research team cooperated and two different versions of the questionnaire were prepared after exploring the literature on best practices in design. Each version has a slightly different approach to providing accessible content while still following the main recommendations such as font, font size, using accessible language a colour, etc. (WCAG, 2022). In Table 2 the points in which the two designs differ from each other can be seen.

Table 2

Design Elements Considered for the Two Preliminary Questionnaire Versions: Design 1 (D1) and Design 2 (D2)

Design Category	Element of the design	Design 1 (D1)	Design 2 (D2)
1. Media	1.1 Visual response alternatives	Yes	Yes
	1.2 Alternate text to images	Yes	Yes
	1.3 Excluding elements that can cause visual distraction when possible	No	Yes
2. Text	2.1 Font size of 16pt or bigger	Yes	No
	2.2 Line spacing more than 1.5	Yes	No
	2.3 Recommended font type (Arial, Helvetica, Tahoma, Calibri, Verdana and Times New Roman)	Yes	Yes
	2.4 Descriptive and clear pages and titles for the instructions	Yes	Yes
	2.6 Break up large amount of text into smaller paragraphs	Yes	Yes

links	3.1 Provide additional explanations for clarity	Yes	Yes
	3.2 Gestalt principles (Proximity & Similarity)	Yes	Yes
	3.3 Radio buttons instead of typical table	Yes	No
	3.4 Conveying information through multiple medium	Yes	Yes
	3.5 Use headings	Yes	No
4. Colors	4.1 Accessible colours	No	Yes
	4.2 Contrasting colours	Yes	Yes
5. Assistive technologies	5.1 Identify document language	Yes	Yes
	5.2 Compatible with assistive technology	Yes	Yes
	5.3 Use tables wisely	Yes	No
	5.4 Export Word into barrier free PDF	Yes	Yes

One of the main elements that diverge in their implementation in each design is the approach to the colour scheme and the interpretation of ‘minimal elements of distraction.’ While the first design (D1) was created using a more colourful scheme (blue, black, and white), the second design (D2) makes use of a different colour scheme (grey, black, and white) with high contrast. The latter design uses the Gestalt principles of proximity and figure-ground (Wagemans et al., 2012) to ease the reader in making the differences between questions and answering options, as well as to separate visually distinct types of information presented (Figure 1).

Figure 1

Implementation of Gestalt Principles of Proximity and Figure-ground in D2

Section one instructions:

In this section you are given three sentences. Choose only one of the possible answers on how much you disagree or agree with the sentence. Put **X** or **O** over the number of your choice.

This is an exemplary sentence with an answer:

This sentence is written in English.

This sentence is written in English.

1. I am familiar with chatbots or other conversational agents.

2. I know how chatbots work.

Note. *The figure shows how part of the instructions for filling in the questionnaire and the first two items from the scale were presented in D2 based on Gestalt principles. The principle of proximity is implemented through borders and the positioning of verbal and numeral answering options for the scale. The principle of figure-ground is implemented by introducing different background colours to separate the main part of the scale and additional information, as well as choosing an appropriate background colour to highlight the presented text.

** This figure does not accurately represent the elements' size in the prototype.

The Gestalt principle of proximity has been implemented via the use of surrounding borders for each question to suggest the necessity of elements to be viewed together. The principle of figure-ground is evoked by increased contrast between text and background – in this

case, done by using bold text on simple background for the questions. In the instructions section as well as by introducing distinct colours to signify different elements of the scale (instructions and items).

The second significant difference between the two designs lies in the choices made for implementing element 1.1 - Visual response alternatives. Both make use of a 5-point Likert scale ('Strongly Disagree' to 'Strongly Agree'), as it is the scale used in the original version of the BUS-11 (Borsci et al., 2022). However, to create a design that can be more inclusive toward people with various levels of disability, it is advised to provide more than one primary medium for the context of information presented (Brewer, 2018; Goegan et al., 2018a). In D1 the solution to this issue was the design choice of including smiles (Figure 2) as an additional representation for each answering option. The smileys were coloured in light grey to retain neutrality and not bias participants' choices (Asmal et al., 2022).

Figure 2

Alternative Answering Options Provided in D1

1. The chatbot function was easily detectable. 🔍

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5
☹️	😞	😐	😊	😄
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

Note. The figure shows what answering options were provided in D1. The design makes use of smileys and a combination of numerical and verbal queues to indicate the strength and direction of the potential response.

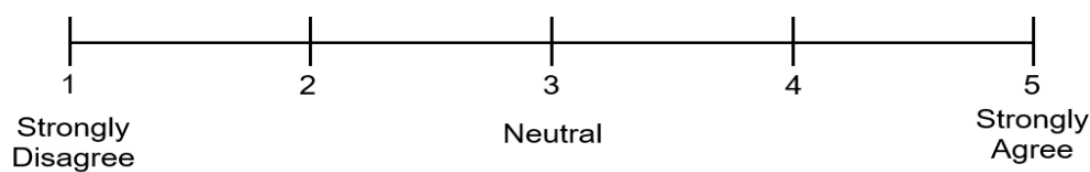
In D2 instead of smileys the answering options were accompanied by a graduated Visual Analogue scale (VAS) and numbers (from 1 to 5, 1 being associated with 'Strongly Disagree' and respectively 5 with 'Strongly Agree'). This type of scale is often used in medical research for measuring the intensity of a symptom and has proved to be a reliable indicator to assess

concepts that involve an underlying continuum (Shiina, 2021). The VAS was measured to be 10 centimetres (about the length of the long edge of a credit card), with each answering choice spaced equally (Figure 3).

Figure 3

Alternative Answering Options Provided in D2

1. I am familiar with chatbots or other conversational agents.




Note. This figure might not accurately represent the true length of the used VAS in the designed scale.

When designing for accessibility it is important that the used items are understandable and easy to follow (Goegan et al., 2018a). The language used in both versions made use of accessible terms for the provided instructions. If a question contains terms that are hard to understand for the participants, they can feel discouraged and withdraw their consent to participate. Additionally, questions should not influence the answers of participants with their wording (Goegan et al., 2018a; Rowley, 2014). So far, no such issues have been found in the BUS-11 (Borsci, 2021; Borsci et al., 2022; Lopez, 2021), but the sample used for validation does not represent accurately people with disabilities. The items of the original BUS-11 were deemed to contain only accessible language by the research team; therefore, the wording was not changed. This provides the opportunity to validate the scale language with people from the target group of this design, namely people who identify as having a disability that could influence their performance when using the scale.

To fully mimic the condition under the designed questionnaires will be used, it is needed to provide a mockup task. The BUS-11 has been tested with multiple chatbots so far (Borsci, 2021). For this study, the Zoom chatbot was deemed appropriate to use. Zoom is a video and conferencing app that gained popularity during the COVID-19 lockdowns (*Video Conferencing, Cloud Phone, Webinars, Chat, Virtual Events - Zoom*, n.d.). Used in education, business and leisure, this app has a great variety of users and therefore it is likely for testes to be familiar with

the environment and therefore better grasp the task provided. An imaginary scenario that participants would have to follow was created. Providing context and not simple instructions for the users to follow has the benefit of freedom of interpretation and therefore resembles the natural circumstances under which people would generally seek help from a chatbot. The task was formulated as follows:

Imagine you are in a Zoom meeting with a friend, preparing for your work together. You start experimenting with the settings in Zoom and remember seeing different backgrounds on other people when they use Zoom. Unfortunately, your friend also does not know how to change the background. So, you decide to get help on this via the Zoom website. Now, your task is to find the chatbot and ask for help. Please open the official Zoom website and look for the chatbot function. Often it is a chat symbol popping up in the corner like the one you can see here . Then, try finding needed information through the suggestions that the chatbot provides you with. After finding the video with the instruction, your task is finished, and you can come back here.

Please follow the link to conduct the task: <https://zoom.us/>

Remember that we are interested in your experience with the chatbot, if for any reason you cannot achieve the goal in a reasonable amount of time, please simply come back here once that you gain enough knowledge to assess the quality of the chatbot.

Method

Study design

To establish which design is more favourable and has better chances of being useful to a wider range of people, testing with participants is needed to validate the design choices. It is recommended when validating a new scale design to use a mixture of qualitative and quantitative methods, depending on the design stage (Carpenter, 2018; Zhou et al., 2019). A focus group setting was chosen as the most appropriate. This method has the benefit of small sample size with strong face validity, thus it provides a convenient opportunity to obtain qualitative data from representatives of the target group.

The first phase of this study will consist of creating two preliminary designs that will be tested in the setting of a focus group with representatives from the target group, namely people

with distinct types and levels of disabilities. The two designs will be created by the research team associated with this study. Once the designs are ready and the main points of exploration are noted a focus group will be organized.

As a result of the focus group, a final version of the BUS-11 will be created. This version will incorporate features from both D1 and D2. Participants will be asked for feedback on the design choices that were made – layout, colour, font, font size, language, answering options and perceived accessibility. Special attention will be given to points which differ in the two designs, such as the different provided answering options.

The focus group will optimally consist of 2 to 5 representatives of the target group and if necessary, their legal guardians. A plan that consists of five phases for the focus group was devised (see Appendix A). The planned length is between 40 minutes and 1 hour. This focus group makes use of both within-subject and between-subject design as each participant will be asked to compare the two different versions according to their own preferences. Then the participants will be encouraged to verbalize the positives and negatives of each design.

Participants

Ethical approval from the University of Twente Ethics committee was obtained before participants were recruited. Participants for the focus group were recruited with the help of School of Open Minds, CA, USA. Open invitation was sent out to students and their guardians and those who expressed interest in participation were invited to join the focus group with the aim being to have around 3 to 5 participants.

The total number of participants recruited was 2, both of whom were male. One participant was 14 at the time of the focus group, and the other was 21 years old. Consent was obtained from legal representatives of both the participants. Both of the participants have been previously diagnosed with autism spectrum disorder (ASD).

Materials

A consent form, also approved by the University of Twente Ethics committee for usage in research was adjusted to fit the specific needs of this focus group and was presented to the participants (see Appendix B).

A mock-up task was prepared to help participants understand better the content of the questionnaires. One task was formulated, and the same formulation was given with both versions to participants (see Appendix C).

Two different accessible design versions were created to explore how different elements are perceived by representatives of the potential end-users. The first design D1 contains demographic questions and the adapted BUS-11 version (see Appendix D). The other design (D2) does not contain demographic questions and has a different interpretation of the core design principles discussed above (see Appendix E).

Additionally, a video displaying a task performed on a chatbot selected by the research team was prepared. The video fulfils the purpose of introducing participants to the idea and core functionalities of chatbots briefly to ensure they are aware of how to continue with the task. The video was not pre-recorded specifically for the focus group, it was a section from a video on the topic of ‘What are chatbots?’ uploaded on YouTube (GCFLearnFree.org, n.d.). The segment selected was 45 seconds long and started from the beginning of the video.

Procedure

The focus group was conducted at the Open Minds School, Silicon Valley, CA. Since both primary researchers were unable to physically be present during the focus group, a detailed protocol of action was given to a representative who volunteered to play the role of representative researched from the school who kindly agreed to help by playing the role of facilitator. The detailed protocol can be seen in Appendix A.

First, participants were briefed on the structure of the focus group and what they can expect from the present research associate. The participants were asked to pay attention to a short introductory video on the topic of chatbots. After the segment was shown, participants were asked if they wanted additional clarification on chatbots and if not, the prepared task with the Zoom chatbot was given to them. Since the main topic of interest of this focus group is not the performance with the chatbot itself, participants were informed they do not need to complete the task, but they are asked to try it to grasp the purpose of the following questionnaire. If participants failed in finishing the task in a reasonable time, they were prompted to continue to the next phase of the focus group.

Consequently, participants were given one of the redesigned versions (D1 or D2) and were asked to fill in the questionnaire while paying attention to what they liked and dislike in the given design. Additionally, participants were prompted to think aloud while exploring the questionnaire to identify areas of the design that had unforeseen drawbacks. The same procedure

was repeated with the second design version. The order of the designs to be presented was not predetermined but was the same for all participants.

Before ending the focus group an open discussion was encouraged between the participants and the facilitator. In the Protocol (see Appendix A) the specific prompts used by the present supervisor can be seen. Those prompts were drafted on the base of concrete design elements that could be potential weak or robust design points. One example of such a prompt is ‘Is the flow of the questionnaire easy to follow?’.

Data Analysis

The recording of the focus groups was transcribed, and the names and any personal information of the participants were removed. The full transcript can be seen in Appendix F. The two members of the research team individually reviewed the transcript and then discussed the results. Three things were inspected: information about what participants liked about each design, information about issues regarding design elements and textual information, and potential helpful insights into what can be improved in the design.

Results

The transcript of the focus group was used to examine the preferences of the participants. Each of the main design categories in Table 2 was given specific attention and was coded as a separate point of interest in the focus group (Table 3). Both participants in the focus group indicated they have an overall preference for D2. Since some of the participants used communication partners in the focus group, this was extrapolated from different segments of the transcript (see Appendix F).

Table 3

Participants' Feedback on Specific Design Points from Each Questionnaire Version

Design Element	Participant 1 issues		Participant 2 issues	
	D1	D2	D1	D2
Font size	1	1	1	1

Use of colour	0	0	0	Likes the colour
Length of questionnaire	0	1	0	0
Comprehension of Likert scale for answering options	Answering options were understandable	Answering options were understandable	Answering options were understandable	Answering options were understandable
Multiple media used for the representation of answering options	Affinity towards smiles	0	Affinity towards smiles	0
	Comprehensible	Comprehensible	1	Comprehensible
Comprehension of items				
Total issues declared	1/6	2/6	2/6	1/6

Note. In this table, if a participant voiced issues with one of the listed elements, this is noted by the number 1. In case the participant showed no specific attitude, this is indicated by the number 0. Quotes are provided for positive attitudes towards items.

Implications of focus group

Based on the insights obtained the final design of BUS-A was proposed. The feedback from the focus group was carefully reviewed and where appropriate incorporated by either improving upon elements that were already part of one of the two designs, combining elements and creating new visual layouts or introducing new elements as a response to participant's comments. Together, the focus group and the preceding systematic literature review supplied sufficient information to answer RQ1: 'What are design elements that can be implemented to improve the accessibility of a scale design based on scientific literature and design recommendations?'. In the following paragraphs, significant design choices that were reconsidered and changed to reach the final BUS-A version will be discussed.

As both participants expressed a negative attitude towards the size of the font in the two questionnaires. According to the WCAG accessibility standards, font size should be 16pt or above (WCAG 2.0 Conformance Levels - UCOP, n.d.; WCAG 2 Documents - Web Accessibility Initiative (WAI) - W3C, n.d.) and for the initial design, the minimal acceptable size was chosen. However, upon reviewing the transcript it became evident that a bigger text would have been preferred. After reviewing the implemented elements, based on both literature and the

participants' feedback, the bigger font size and more space between the elements suggest an improvement in the overall experience of using the scale. As an outcome, the text size was increased to 18pt.

The participants unanimously agreed that there are no issues present in the formulation of the 11 items of the scale. The wording was found understandable and unproblematic by both. This goes to show that the items do not include inaccessible language and can be used in the BUS-A. Therefore, the second research question RQ2: 'Does the current item formulation of the BUS-11 comply with accessibility requirements?' can be answered. The current item formulation of the scale does not pose a threat to the accessibility of the BUS, hence there is no need for adjustments and the items can be directly used in the new proposed design. The same can be said about the task formulation since neither participant expressed any difficulties with understanding the provided scenario. However, one participant did express difficulties with answering two of the questions in D1, which as described by them were '... hard to answer...' (see Appendix F). When prompted to elaborate the participant indicated the questions are 'I am familiar with chatbots.' and 'I know how chatbots work.' which are not part of the main 11 items included in the scale. Discussing the matter with the other researcher on the team, it was decided to omit the section from the finalized design, as it does not influence any of the five main underlying factors of the BUS-11 and does not change the integrity of the scale.

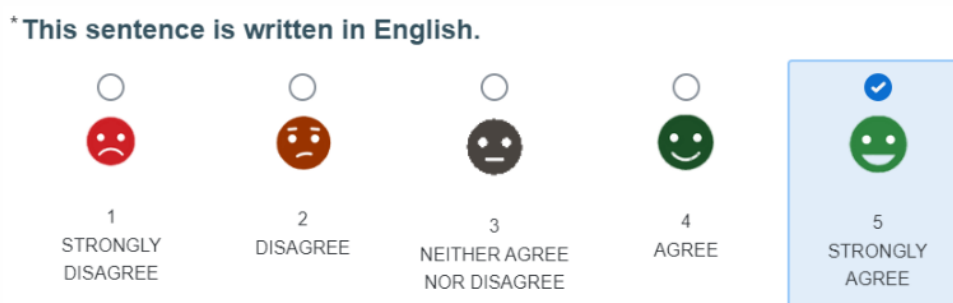
Another design point that both participants showed a strong preference for was the inclusion of smiles in D2 as an addition to the answering options. With both participants liking the element of smiles, they were included in the final design to help convey the meaning of the 5-point Likert scale via an alternative medium. Because both participants also enjoyed the inclusion of colours in the same design, as opposed to having no opinion on the grayscale colour scheme of D1, it was decided to further improve the ways the smiles communicate the meaning of each answering option by using universally recognized colour scheme to match with the assigned level of satisfaction by each face. A combination of shades of red, green, and grey is widely used to call upon association with positive / higher and negative / lower satisfaction (Northway et al., 2015). All the shades used are approved as accessible options according to the WCAG standards. An exemplary item from the finalized version of the scale after implementing the changes mentioned so far can be seen in Figure 4. Additionally, the options were supported

by numbers in growing order and the verbal quest to indicate the range of choices ranging from ‘Strongly Disagree’ on the left to ‘Strongly Agree’ on the right.

Figure 4

Item Used as an Example for Participants with the BUS-A Design.

Here you can see an example question with an answer:



Note. The figure shows an example of the answering options in the final BUS-A design. Each option is presented with a combination of a smiley, numeral indication for relative position and verbal queues that aid interpretation.

The finalized design was prepared for use in both paper and online format. In both versions, one item per page will be presented to participants to account for the increase in size and provide adequate space for the answering options to be displayed. The whole paper version of the questionnaire can be seen in the supplementary materials (see Appendix G). A digital version of the same design was adapted to Qualtrics (Qualtrics, 2021). The BUS-A includes a section with demographic questions, a sample task with the Zoom chatbot and the 11 items of the scale.

To ensure that this design can be used in further research involving the BUS- A the scientific validity and reliability of the scale need to be confirmed.

Phase Two: Validation of the BUS-A

The second phase of the study aims at validating the psychometric properties of the BUS-A as a scale and establishing if it is comparable to the BUS-11. This can be achieved through quantitative methods such as survey data collection (Dalati & Marx Gómez, 2018). Since a new design is being proposed, it is central to establish the legitimacy of its relation to the original scale it is being derived from (Zhou et al., 2019).

Method

Study Design

The BUS-A design was compared against the original BUS-11 developed by Borsci et al. (Borsci et al., 2022; Lopez, 2021). The design of the study was compromised on two conditions: control and experimental. Before a participant was assigned to one of the two conditions, which was done in a semi-randomized manner, they were presented with the same initial questions related to their age and gender, as well as a question asking about any disability that might affect their experience with the chatbot. Participants were divided into two conditions: control and experimental.

In the control condition, the previous format of the BUS-11 was preserved as used by Borsci et al. The participants were asked first for their level of proficiency in English, after which they were presented with three out of five tasks on different chatbots. The Zoom chatbot was presented to each participant regardless of which of the two groups they were randomly assigned to. Two other chatbots in each condition were given to collect data for another researcher with the BUS-11. The additional four chatbots were provided by Mustafa Taha, a bachelor's student whose work also involves using the BUS-11. As his study makes use of the original format of the BUS-11, an agreement to cooperate and collect data using comparable questionnaires to provide a larger sample was made.

Each task was followed immediately by a control question on whether the participant completed the task. Afterwards, the 11 items for the BUS-11 were adjusted for the specific chatbot. Participants were also inquired to fill in a condensed version of the UMUX – Lite (Lewis & Sauro, 2020) scale consisting of three questions related to the overall performance of the product, in this case, the corresponding chatbot.

In the experimental condition, participants used the BUS-A after the interaction with a chatbot. All participants who gave ‘Yes’ as a response to the question of whether they have any disability that could affect their performance on the scale were assigned to this condition. To balance out the number of participants across conditions, some of the participants who did not indicate any disability were assigned to the experimental group as well.

In the experimental group, only one chatbot was given per participant, namely, the Zoom chatbot that has been previously used in the focus group was presented with the same task

formulation that can be seen in Appendix C. Participants in this condition were only presented with the 11 items included in the BUS-A without the UMUX- Lite follow up questions.

Participants

The total number of participants across the two conditions after including the data collected from another researcher on the Zoom chatbot by using the BUS-11 was 116 participants ages ranging from 15 to 60 years old ($M = 24.87$, $SD = 8.28$). All participants gave their consent for the data collected to be used in this research after being informed of its purpose.

In the control group, there were 90 participants, out of which 52 identified as female, 36 identified as male, one participant indicated they identify as non-binary, and one participant as genderfluid. The average age of participants in this group was $M = 24.19$, $SD = 7.63$, with the youngest participant being 15 and the oldest – 54.

For the experimental group 26 participants were recruited, out of which 18 identified as female and 8 as male. The age range for this condition was between 21 and 60 years, $M = 27.31$, $SD = 10.07$. Out of the participants in the experimental condition, five indicated they have some form of disability that could influence their performance with the chatbot. Out of them, two people used more than one term provided to describe their disability. Both reported having an unseen disability and sensory disability, with one indicating developmental disability as well and the other indicating a mental or emotional disability. The other three people each used one term to describe the nature of their disability, with one reporting mental or emotional disability, another reporting a physical disability and the last person reporting a sensory disability.

Before recruiting participants, permission to proceed to from the Ethics committee was obtained. Each participant was asked for consent to use their data was obtained after briefing them on the purpose of the research. The study was published on the University of Twente SONA system where students can take part in study in exchange for credits. Next to that, the study was advertised via personal social media and Linked-In, as well as several survey-exchange groups on Facebook and Reddit.

Materials

The BUS-A and the BUS-11 designs were used to obtain information on the user's satisfaction with the products, in the case of this study – with various chatbots. Both versions make use of a modified 5-point Likert scale, ranging from '1 – Strongly Disagree' to '5 – Strongly Agree'. The two designs have the same 11 items distributed over 5 underlying factors

as described in Table 1: Factor 1 - Perceived accessibility to chatbot functions; Factor 2 - Perceived quality of chatbot; Factor 3 - Perceived quality of conversation and information provided; Factor 4 - Perceived privacy and security; Factor 5 - Time response.

Participants received the tasks and the respective scale via Qualtrics survey link. The scales were divided over two conditions: control condition that uses the BUS-11 and experimental condition that contains the BUS-A.

The task given was the same in both variations. As the Zoom chatbot did not show any issues in the previous stage of the study, it was used in this stage as well. Participants were presented with a scenario and asked to imagine they were a part of it. The goal of the task was to find information on changing their background in a virtual meeting. In Supplemental C the concrete task with a link directed to the chatbot included can be found.

In addition to the scales a form for consent to participate was prepared. In this form the purpose of the study, along with general information about how the collected data will be handled and some requirements for participation that must be met were presented to potential participants. The formulation of the text can be seen in Appendix H.

Procedure

Upon opening the questionnaire link, information about the aim of the study and the requirements for participation were presented. If a person did not give their consent to participate or indicated not being eligible based on the provided requirements for participation, they were redirected to the end of the survey and no data was collected from that entry. All participants who did fit the criteria and gave their consent were first introduced to the demographics section of the questionnaire. In this section questions about the age and gender of the participants were asked. A question about perceived disability was also posed. After finishing the demographic section, participants were set to evenly distribute across the two conditions, which was automated with Qualtrics. An exception was made for participants who indicated having a disability on the demographic's questionnaire. They were automatically redirected toward the experimental condition that contains the BUS-A version. After being allocated to a condition, respondents were presented with the task and a link to the chatbot. This was followed by the 11 items of the scale with either the BUS-A design or the BUS-11. In the control condition, participants were asked to perform tasks on two more chatbots and this data was collected and given to Mustafa Taha and discarded from any datasets related to this current study. After

completion, all participants were presented with an opportunity to leave feedback and thanked for their participation.

The study has a between-subjects experimental design with the independent variable being whether they received the regular or BUS-A version of the scale and the dependent variable being the measured user satisfaction. Since the internal structure of the scale is important for the aim of the study, only entrees that had all 11 items answered were considered complete.

Data analysis

Data was collected over the period between 09/04/2022 and 11/05/2022. All responses outside of this data frame were considered invalid. After exporting the data from the Qualtrics platform, it was transformed before proceeding with further testing. Participants who did not consent to have their data recorded or indicated they do not comply with the criteria for participation were excluded from the data set. Questionnaires that were not completed, meaning they have less than the full 11 items filled in, were deleted as well. Additionally, the two conditions were separated into two different data sets labelled 'Control' and 'Experimental.' The collected data from the two conditions were treated as ordinal data. Therefore, to prepare it for analysis each response was standardized. With the highest obtainable score being considered as 100% satisfaction with the chatbot, each questionnaire entry was standardized based on the percentage of points given out of the highest possible.

First, to ensure the quality of the data, deviations from the criteria for inclusion will be removed, as well as any data which has errors in recording and could have a negative impact on the analysis. Exploratory analysis for initial trends in the responses will be performed to check for abnormalities in the data. To answer the RQ3: 'Is the factorial structure of the BUS preserved in the BUS-A, based on CFI value of 0.90 or greater and RMSEA value of less than 0.08, and is the new design a reliable scale to measure user satisfaction based on Cronbach's Alpha value equal to 0.7 or greater?' a mix of inferential and exploratory analyses will be performed on both conditions. Comparative analyses between the two conditions, such as a two-tailed independent sample test will be performed to search for significant differences between the two conditions.

The reliability on both scales will be tested, with the expectation that the BUS-11 will show results consistent with previous (Cronbach's Alpha of $\alpha = .90$) and the BUS-A will produce $\alpha > .70$ to indicate internal reliability. The underlying factorial structure of the two

versions will be tested by performing a Confirmatory Factor Analysis (CFA) with the established 5 – factors model of the BUS-11 on both. This type of analysis is used to confirm the existence of the already established factor model in the collected data and will be performed on the BUS-11 to confirm there are no deviations from the expected compliance with the factorial structure first, and after that it will be performed on the BUS-A to confirm the existence of the same underlying structure. To confirm the factor structure of the BUS-A the following criteria will be adhered to: RMSEA value of less than 0.08 is considered to be an acceptable fit; CFI value of 0.90 or greater.

Additional visualizations of data such as distributions, Q-Q plots and histograms of the scores will be used to compare the two conditions and infer if there are non-explainable differences that need to be addressed.

Results

Before applying the exclusion criteria to the data set a total of 120 recorded responses were counted. After removing data that is not complete and does not meet the criteria a total of 79 entries remain. Additional data conformant to all of the quality criteria and collected on the Zoom chatbot for the control condition was obtained from Mustafa Taha. After this the total number of participants across the two conditions was 116. From them 90 were in the control group and 26 in the experimental condition.

For each of the two conditions a variable ‘Percentage’ was created to indicate the total percentage of user satisfaction with the given chatbot. This variable was operationalized as the ratio of the actual score obtained from this data unit to the maximum possible score for customer satisfaction to normalize the data. Next to that, the five underlying factors were operationalized by combining the associated item scores and obtaining factor specific standardized scores (Table 4).

Table 4

Factors of the Assumed Underlying Model with Corresponding Items

Underlying factor	Item
Factor 1	Item 1, Item 2
Factor 2	Item 3, Item 4, Item 5
Factor 3	Item 6, Item 7, Item 8, Item 9

Factor 4

Item 10

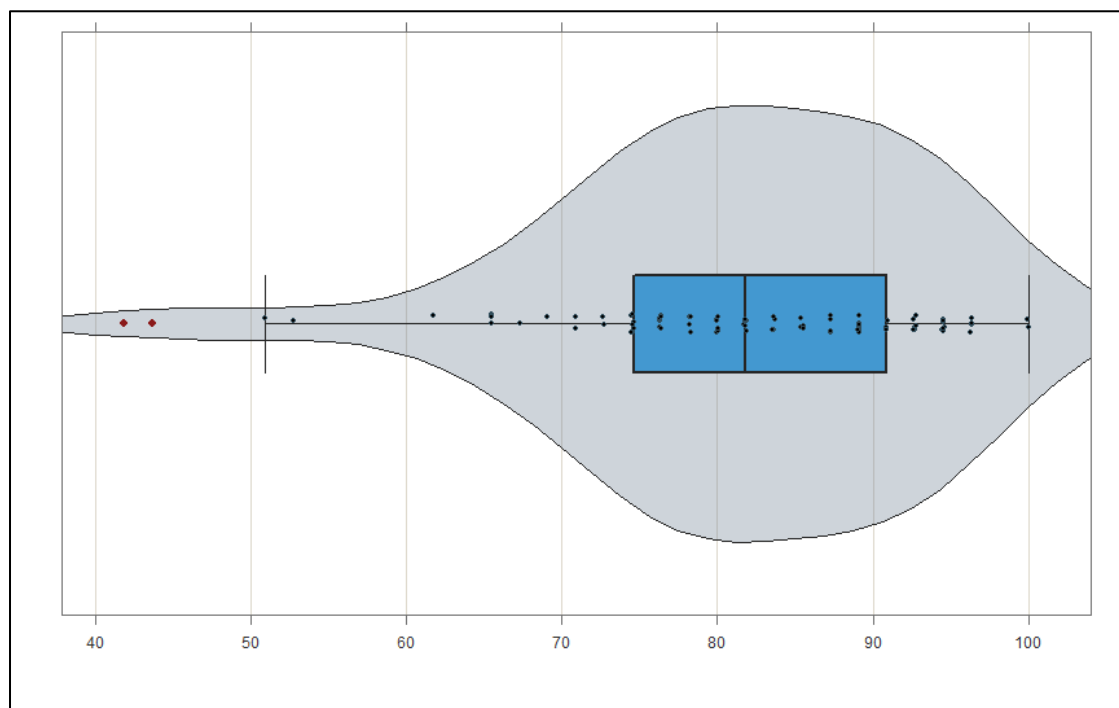
Factor 5

Item 11

The distribution of scores was checked in both conditions with a violin plot for each condition for rough comparison (Figure 5 and Figure 6). Afterwards, each set of data was additionally analysed for a more detailed understanding of the normality. In the control condition, the skewness of the total user satisfaction was found to be -0.81, indicating that the data is skewed to the left. The kurtosis of this group was found to be 0.81, which can be seen as more heavily tailed in comparison to a normal distribution. Respectively, the skewness for the experimental group was -0.01, which is close to a normal distribution with a kurtosis of -0.5656. The data in the control condition therefore was treated as normally distributed, but the data in the experimental condition does not fit the requirements for normality and shows tendencies resembling positively skewed data, albeit not extreme ones.

Figure 5

Violin Plot of Recorded Scores in the Control Condition

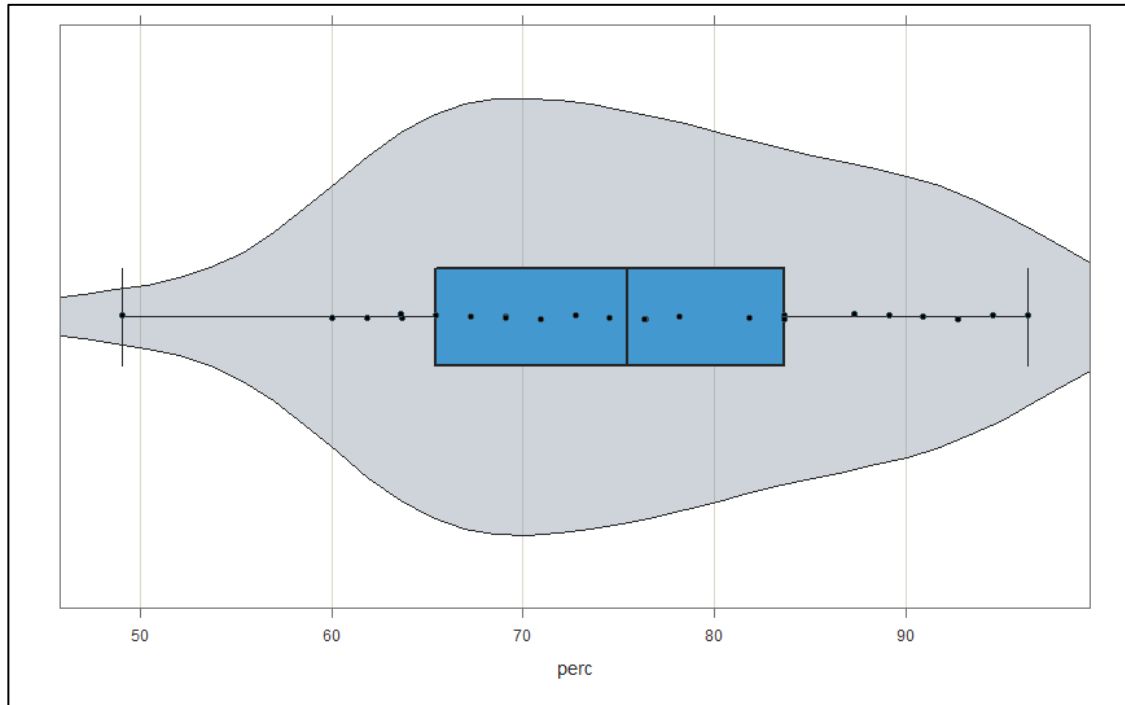


Note. This figure depicts the distribution of the total customer satisfaction score in percentages for the control group with number of data units recorded $N = 90$ ($M = 81.41$, $SD = 11.72$). The lowest value recorded in this condition

was MIN = 40.82 and the highest MAX=100. In the control condition, two outliers below the first quartile were recorded.

Figure 6

Violin Plot of Recorded Scores in the Experimental Condition

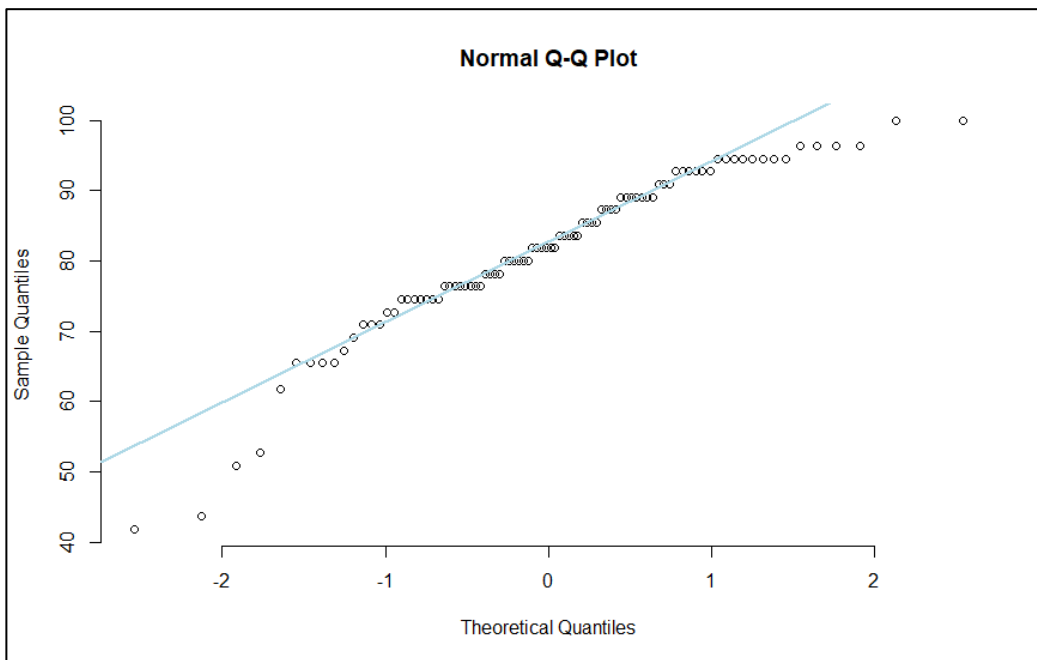


Note. This figure depicts the distribution of the total customer satisfaction score in percentages for the experimental group of this study with $N = 26$ data units ($M = 75.45$, $SD = 12.05$). The lowest recorded value was $MIN = 49.90$ and respectively the highest was $MAX = 96.36$.

In the next step of analysis normality of the data was her explored with q-q plots in both conditions. In the control condition, the scores show overall normal distribution, with tails on both sides (Figure 7). The data collected in the experimental condition similarly shows light-tailed normal distribution (Figure 8).

Figure 7

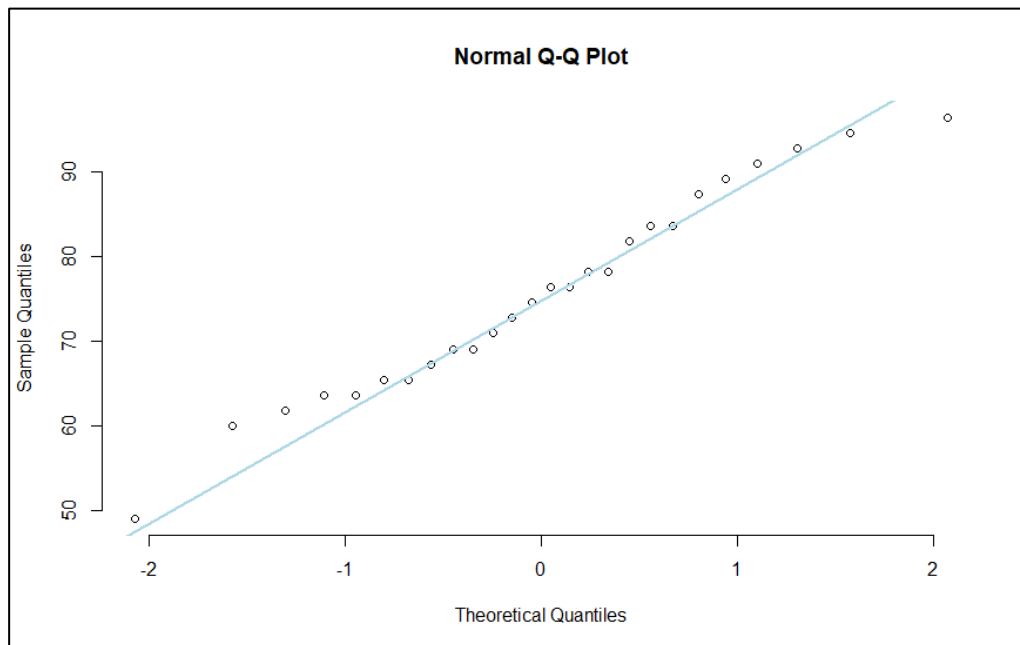
Distribution of the Recorded Scores in the Control Condition



Note. The sample used for this visualization is the control group participants, N = 90.

Figure 8

Distribution of the Recorded Scores in the Experimental Condition



Note. The sample used for visualization is the experimental group, N = 26

The normality of each factor was analysed with the Shapiro-Wilk test with the hypothesis that each factor is normally distributed in the population (Table 5). Across the two conditions, the null hypothesis that the population distribution is normal was rejected for all the factors in the control condition based on $p < 0.05$. However, for the experimental condition, the hypothesis was rejected for all factors but for Factor 3. For this factor based on $p > .05$, the hypothesis was confirmed, meaning the sample for this factor comes from a normal distribution.

Table 5

Shapiro – Wilk Test of Normality for Each Factor

	Control Condition			Experimental Condition		
	Statistic	df	Sig.	Statistic	df	Sig.
Factor 1	0.78	89	<.01	0.82	25	<.01
Factor 2	0.88	89	<.01	0.91	25	.02
Factor 3	0.91	89	<.01	0.96	25	.42
Factor 4	0.90	89	<.01	0.86	25	<.01
Factor 5	0.65	89	<.01	0.68	25	<.01

Differences in Mean Scores Between Conditions

The 90 participants in the control condition ($M = 81.41$, $SD = 11.72$) compared to the 26 participants in the experimental condition ($M = 75.45$, $SD = 12.05$) demonstrated significantly higher satisfaction scores, $t(114) = -56.42$, $p < .05$. Each of the five factors was separately analysed in both conditions to see if there is an effect of the test condition on any of them. Table 6 shows the summarized findings for each factor. Factor 3 was the only one that showed significant differences between the scores recorded in the control and experimental condition, based on $p < .05$.

Table 6

Two Tailed Independent t-Test Between Each Factor for the Two Conditions

	Welch Two Sample t-test		
	<i>t</i>	<i>DF</i>	<i>p</i>
Factor 1	-0.23	114	.81
Factor 2*	1.87	33.43	.07
Factor 3	2.85	114	.01
Factor 4	1.49	114	.13
Factor 5	-0.13	114	.89

Note. In the table the factors that showed significant differences in variance of scores recorded ($p < .05$) between the two conditions are marked with *. Those factors were tested under the assumption of unequal variance.

For the participants in the experimental condition, a separate analysis was performed to compare the satisfaction scores between participants who did not report a disability ($N = 20$) and participants who did ($N = 6$). There was no significant effect of reporting disability on the satisfaction scores, despite participants who did not report a disability ($M = 66.81$, $SD = 10.91$) having overall higher scores than those who did ($M = 63.91$, $SD = 16.11$).

Additionally, the satisfaction scores of the 90 participants in the control condition ($M = 81.41$, $SD = 11.72$) were compared to the 20 participants from the experimental condition who did not indicate a disability ($M = 66.81$, $SD = 10.91$). The participants in the control condition reported significantly higher scores compared to this subsection of participants in the experimental condition, $t(119) = 1.94$, $p = .054$.

Reliability Analysis

Both conditions consisted of 11 identical items. The BUS-11 used in the control condition (11 items; $\alpha = .86$) and the BUS-A from the experimental condition (11 items; $\alpha = .83$) were both found to be highly reliable. In the BUS-A factors one (Item 1 and Item 2), two (Item 3, Item 4, and Item 5), and three (Item 6, Item 7, Item 8, and Item 9) showed Cronbach's Alpha values of .63, .88, and .83 respectively.

Confirmatory Factor Analysis on a Five-Factor Structure

On each condition, CFA was performed separately to confirm the expected underlying structure. was performed for each of the conditions. In Table 7 information about the goodness-of-fit indicators for each group can be found. Based on the information presented, the model has shown to be a good fit in both of the conditions. According to the pre-established criteria the control condition meets the requirements for confirming the fit of the proposed factorial structure, the BUS-A fails to meet the required values of less than 0.08 for RMSEA and value of less than 0.90 for CFI.

Table 7

Goodness – of – Fit Indicators

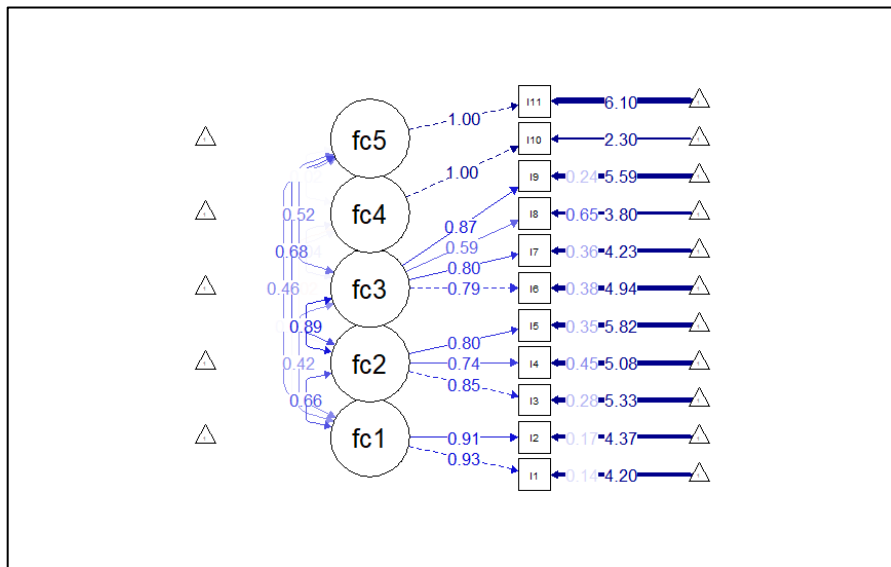
	Indicator				
	Chi square	df	CFI	RMSEA	SMSR
Control condition (BUS-11)	66.31*	36	0.945	0.097	0.053
Experimental Condition (BUS-A)	74.29*	36	0.759	0.202	0.131

Note. The chi-square values that are marked with * are statistically significant results based on $p < .05$.

The underlying factorial structure was plotted for convenience in order to display all the factor loadings and covariances between the distinct factors. The factor loadings were visualized for both the control and the experimental condition. The analysis for the control condition with BUS-11 confirms the underlying factorial structure (Figure 9). However, in the experimental condition, there is an inverse relationship between Factor 4 and Factor 5 (Figure 10). Based on the requirements set with the third research question RQ3: ‘Is the factorial structure of the BUS preserved in the BUS-A, based on CFI value of 0.90 or greater and RMSEA value of less than 0.08, and is the new design a reliable scale to measure user satisfaction based on Cronbach’s Alpha value equal to 0.7 or greater?’, the BUS-A does not conform to the expected factorial structure, but has a satisfactory Cronbach’s Alpha value of $\alpha = .83$.

Figure 9

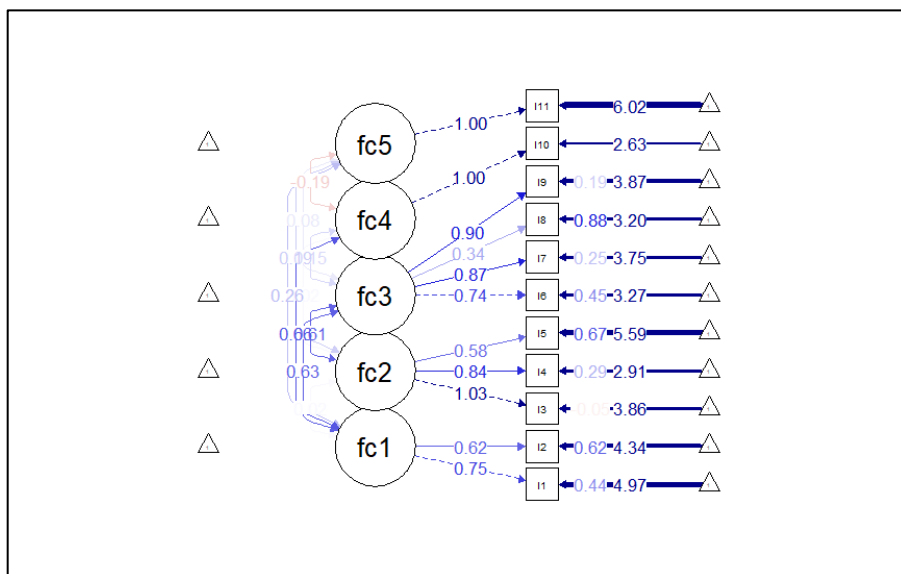
CFA Visual Representation for the Control Condition



Note. The figure depicts the factor loadings of each individual factor of the used model and the variance each explains.

Figure 10

CFA Visual Representation for the Experimental Condition



Note. The figure depicts the factor loadings of each individual factor of the used model and the variance each explains.

Discussion

In this study the goal was to develop and propose a new version of the BUS-11 that could be used in the future as a substitute of the current scale design. The new scale, BUS-A is designed in a way that is compliant with accessibility requirements for various content medium, be it in physical or in digital form (Lim et al., 2021; Ribera et al., 2009; Van Selm & Jankowski, 2006).

This design was conceptualized and confirmed in the first phase of the study with a focus group consisting of two persons, both of whom were diagnosed with a disability. Two out of the three research questions in this work are associated with the first phase. The first and second research questions were answered after a review of existing literature on the topic of designing for accessibility and qualitative testing of two prototype versions for an improved design. Conformance to the W3C Success Criteria was a major prerequisite for the initially proposed design version (W3C, 2016). The new design elements ensure that the relationship between the contents as well as the contents themselves are programmatically determined. Additionally, alternatives to pictures are provided and all functions are made accessible via a keyboard.

Two preliminary versions were designed and tested in a focus group to get additional input on what features are desirable as a part of an accessible scale. Participants' feedback was used to further validate the design choices that were made. Therefore, the answer to RQ1: 'What are design elements that can be implemented to improve the accessibility of a scale design based on scientific literature and design recommendations?' was the resulting final design for the BUS-A. Key findings from this were the importance of relying on multiple media when presenting any type of information. In line with existing literature on the topic, participants showed an inclination towards designs with more vibrant colours and less convoluted page designs (Dalati & Marx Gómez, 2018; Goegan et al., 2018b). Additionally, the focus group results suggested a mix between VAS and traditional Likert scale being the preferred answering options, which does not oppose previous findings on the usefulness of VAS in the context of self-reported satisfaction (Voutilainen et al., 2016). The final design proposed also makes use of an accessible colour pallet for easier differentiation between elements, which was expected and is in line with earlier research on colour implementation in accessibility (Brewer, 2018), and multiple media for conveying information. It, therefore, aids interpretation and makes use of proper spacing and size of the provided items (Dalati & Marx Gómez, 2018). Since in the stage of item formulation of

the BUS attention was paid to ensure unbiased and easy interpretation (Borsci, 2021), the RQ2: ‘Does the current item formulation of the BUS-11 comply with accessibility requirements?’ was easy to answer. Upon review of the items, there were no indications of potential issues, which was later confirmed with the focus group participants, hence the items were preserved in their original state for the new design.

The proposed design, BUS-A, was then tested on a larger scale to explore if the reliability and validity are consistent with the BUS-11 design and whether the underlying factorial structure is preserved. Both designs were tested with the same task and chatbot to ensure that there are no additional influences on the obtained satisfaction scores. Despite the BUS-A showing internal validity that is satisfactory based on Cronbach’s Alpha significance, it does not confirm the factorial structure of the BUS and therefore RQ3: ‘Is the factorial structure of the BUS preserved in the BUS-A, based on CFI value of 0.90 or greater and RMSEA value of less than 0.08, and is the new design a reliable scale to measure user satisfaction based on Cronbach’s Alpha value equal to 0.7 or greater?’ cannot be answered positively.

Because it is a newly proposed design, there is no pre-existing support of the scale validity, but based on the obtained results, it can be assumed that in the future, upon retesting the BUS-A, the scale can be expected to perform reliably under different conditions. The already existing BUS-11 confirmed its psychometric properties, in line with previous testing (Borsci, 2021). However, when the two scales were tested with the assumed factorial structure of the BUS-11, the BUS-A did not show good fit to the existing model. Additionally, two of the factors showed negative correlation between each other, which was not present in the original scale. Negative correlation between factors could suggest that there might be dependency between the scores of the items included and therefore further testing with a different model or sample is recommended (Bollen, 2002). The novel design also produced a significant difference in overall scores of satisfaction reported when compared to the original design. The proposed design is measuring consistently lower overall satisfaction scores and when each factor was tested separately, only four out of five factors showed no significant difference in mean scores. Taking all of this into account, it cannot be said that the BUS-A measures the same concept as the original scale version and therefore further research on creating an accessible version of the BUS-11 is needed. As an attempt to explain what can be causing the discrepancy between expectations and results, the results collected from the BUS-A were investigated more

thoroughly. There seem to be consistently higher scores of satisfaction measured in the control condition, which can indicate issues with how the BUS-A performs when used by people with different disabilities.

This result does not align with the expectations and there are two potential causes that could provide an explanation. The first one is that the model does not apply to the BUS-A and there is a different factor structure that can be used to better explain the scale and the relation between the concepts it tries to measure. Despite the items of the BUS-A being taken from the original scale design without adaptations, the provided answering scale was changed. Visual analogue scales have been used in the past to reliably measure concepts in psychological research (Physiopedia, 2019; Voutilainen et al., 2016), however, in this scale other elements such as colourful emojis were added and their impact on reported satisfaction needs to be tested separately in a controlled manner.

A second factor that could be influencing the results can be related to the power of the sample used in this study. For performing a CFA the recommended sample power is at least 10 cases for each of the estimated parameters, making the suggested minimal sample size for this study 50 people (Kyriazos, 2018). In the experimental condition, there were only 26 recorded responses, which does not meet this recommendation and therefore is prone to producing unreliable and unsatisfiable results. Further exploration of the model is certainly needed with a larger sample size, but also focusing on a different structure. Moreover, the sample power could explain the abnormality in the correlation between factors seen in the experimental condition. Closer look at the factors that display this shows both have one item each, which combined with the small sample size can exaggerate the results artificially (Bollen, 2002).

Limitations

It is important to acknowledge two setbacks this study encountered throughout the various stages that can be seen as limitations. First, the fact that the finalized design of phase one was created based on a small sample of participants in the focus group needs to be accounted for. The literature review and exploration of best practices in design was done to compensate for that. However, there is always added value in confirming the design choices once more (Asmal et al., 2022; Dalati & Marx Gómez, 2018). If this is to be done in the future, one piece of advice would be to aim for a larger group to re-evaluate the accessible scale with people with diverse types of

disabilities who can provide more insights on what elements would aid them while filling in the scale.

The second possible limitation of this study is the sample of participants involved in phase two. Given the BUS-A is designed with inclusivity in mind, it is essential to validate the design with a large and diverse sample of the possible end-users. Despite attempting to do so, validating the design on a larger scale and reaching greater number of people with disabilities was not feasible for the scope of this study. Increasing the participation and recruiting more people to validate the current findings can perhaps lead to different perspective of the perceived fitness for use of the BUS-A.

Conclusion

When developing new products or services it is vital to provide everyone opportunity to benefit from the given product. Making research accessible is a goal researchers have been tackling for a long time (Goegan et al., 2018b). The new proposed design, BUS-A, is compliant with accessibility standards and has indications of being internally consistent. Moreover, the scale measured similarly the satisfaction for both people with disabilities and people without disabilities. This means it can be looked at as useful insights into accessible research design. Despite not being able to prove retaining the same underlying factors as the BUS-11, the BUS-A can be used as a stepping-stone for optimization of the BUS and for eventually designing an accessible version.

Inclusivity is a topic that should not be ignored in any context, especially in research. Regardless of the research stage – planning, recruiting participants, preparing materials, or collecting data, accommodations and adjustments are possible (Rudolf, D., 2017). Lowering entry barriers to participation is not only possible but is necessary for diversifying the scope of research and therefore leading to more productive, accurate and inclusive results.

References

- Asmal, L., Lamp, G., & Tan, E. J. (2022). Considerations for improving diversity, equity and inclusivity within research designs and teams. *Psychiatry Research*, 307(October 2021), 114295. <https://doi.org/10.1016/j.psychres.2021.114295>
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, 605–634. <https://doi.org/10.1146/annurev.psych.53.100901.135239>
- Borsci, S. (2021). *Testing a scale for perceived usability and user satisfaction in chatbots : Testing the BotScale*.
- Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, 26(1), 95–119. <https://doi.org/10.1007/s00779-021-01582-9>
- Brewer, J. (2018). Exploring paths to a more accessible digital future. *ASSETS 2018 - Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–2. <https://doi.org/10.1145/3234695.3243502>
- Carpenter, S. (2018). Ten Steps in Scale Development and Reporting: A Guide for Researchers. *Communication Methods and Measures*, 12(1), 25–44. <https://doi.org/10.1080/19312458.2017.1396583>
- Carroll, A. (2012). World report on disability. *Irish Medical Journal*, 105(5). <https://doi.org/10.1111/j.1741-1130.2011.00320.x>
- Centre for Disease Control and Prevention. (2017). *Disability and Health Overview | CDC*. <https://www.cdc.gov/ncbddd/disabilityandhealth/disability.html>
- Dalati, S., & Marx Gómez, J. (2018). *Surveys and Questionnaires*. Springer International Publishing. https://doi.org/10.1007/978-3-319-74173-4_10
- DSM-V. (2013). Diagnostic and Statistical Manual of Dsm-5 TM. In *Am Psychiatric Assoc*. [http://repository.poltekkes-kaltim.ac.id/657/1/Diagnostic and statistical manual of mental disorders _ DSM-5 %28 PDFDrive.com %29.pdf](http://repository.poltekkes-kaltim.ac.id/657/1/Diagnostic%20and%20statistical%20manual%20of%20mental%20disorders_%20DSM-5%20PDFDrive.com%29.pdf)
- European accessibility act - Employment, Social Affairs & Inclusion - European Commission*.

- (n.d.). Retrieved March 15, 2022, from <https://ec.europa.eu/social/main.jsp?catId=1202&langId=en>
- GCFLearnFree.org. (n.d.). (289) *What are Chatbots? - YouTube*. Retrieved May 6, 2022, from <https://www.youtube.com/watch?v=pX6zqaEHAdw>
- Gilchrist, K. (2017). *Chatbots expected to cut business costs by \$8 billion by 2022*. CNBC. <https://www.cnbc.com/2017/05/09/chatbots-expected-to-cut-business-costs-by-8-billion-by-2022.html>
- Global Health Workforce Alliance. (2019). What do we mean by availability, accessibility, acceptability and quality (AAAQ) of the health workforce? In *World Health Organisation* (p. 1). World Health Organization. [https://www.who.int/workforcealliance/media/qa/04/en/ ?](https://www.who.int/workforcealliance/media/qa/04/en/)
- Goegan, L. D., Radil, A. I., & Daniels, L. M. (2018a). Accessibility in Questionnaire Research: Integrating Universal Design to Increase the Participation of Individuals With Learning Disabilities. *Learning Disabilities, 16*(2), 177–190.
- Goegan, L. D., Radil, A. I., & Daniels, L. M. (2018b). Accessibility in Questionnaire Research: Integrating Universal Design to Increase the Participation of Individuals With Learning Disabilities. *Learning Disabilities, 16*(2), 177–190.
- International Day of Disabled Persons 2004 | United Nations Enable*. (n.d.). Retrieved May 9, 2022, from <https://www.un.org/development/desa/disabilities/international-day-of-persons-with-disabilities-3-december/international-day-of-disabled-persons-2004-nothing-about-us-without-us.html>
- ISO/IEC JTAG. (2014). *ISO/IEC Guide 71:2014 - Guide for addressing accessibility in standards. 2014*. <https://www.iso.org/obp/ui/#iso:std:iso-iec:guide:71:ed-2:v1:en>
- Kyriazos, T. A. (2018). Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General. *Psychology, 09*(08), 2207–2230. <https://doi.org/10.4236/psych.2018.98126>
- Lewis, J., & Sauro, J. (2020). *Simplifying the UMUX-Lite – MeasuringU*. 3 December. <https://measuringu.com/modified-umux-lite-usefulness-item/>

- Lim, Y., Giacomini, J., & Nickpour, F. (2021). What Is Psychosocially Inclusive Design? A Definition with Constructs. *Design Journal*, 24(1), 5–28. <https://doi.org/10.1080/14606925.2020.1849964>
- Lopez, S. M. K. (2021). *Confirmatory Factor Analysis of a new Satisfaction Scale for Conversational agents and the role of decision-making styles*. University of Twente.
- Northway, R., Howarth, J., & Evans, L. (2015). Participatory research, people with intellectual disabilities and ethical approval: Making reasonable adjustments to enable participation. *Journal of Clinical Nursing*, 24(3–4), 573–581. <https://doi.org/10.1111/jocn.12702>
- Physiopedia. (2019). Visual Analogue Scale - Physiopedia. In *Visual Analogue Scale*. https://www.physio-pedia.com/Visual_Analogue_Scale
- Qualtrics. (2021). Qualtrics XM // The Leading Experience Management Software. In *Qualtrics*. <https://www.qualtrics.com/uk/?rid=ip&prevsite=en&newsite=uk&geo=NL&geomatch=uk>
- Ribera, M., Porrás, M., Boldu, M., Termens, M., Sule, A., & Paris, P. (2009). Web Content Accessibility Guidelines 2.0. *Program*, 43(4), 392–406. <https://doi.org/10.1108/00330330910998048>
- Rowley, J. (2014). Designing and using research questionnaires. *Management Research Review*, 37(3), 308–330. <https://doi.org/10.1108/MRR-02-2013-0027>
- Shiina, K. (2021). Commentary: The Historical Roots of Visual Analog Scale in Psychology as Revealed by Reference Publication Year Spectroscopy. In *Frontiers in Human Neuroscience* (Vol. 15). Frontiers Media S.A. <https://doi.org/10.3389/fnhum.2021.711691>
- United Nations Convention on the Rights of Persons with Disabilities - Employment, Social Affairs & Inclusion - European Commission*. (n.d.). Retrieved March 15, 2022, from <https://ec.europa.eu/social/main.jsp?catId=1138&langId=en>
- Van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. In *Quality and Quantity* (Vol. 40, Issue 3, pp. 435–456). <https://doi.org/10.1007/s11135-005-8081-8>
- Video Conferencing, Cloud Phone, Webinars, Chat, Virtual Events | Zoom*. (n.d.). Retrieved June 15, 2022, from <https://zoom.us/>
- Voutilainen, A., Pitkäaho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2016). How to ask about

patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, 72(4), 946–957. <https://doi.org/10.1111/jan.12875>

W3C. (2016). *Understanding Conformance | Understanding WCAG 2.0*. Understanding Conformance. <https://www.w3.org/TR/UNDERSTANDING-WCAG20/conformance.html#uc-levels-head>

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, 138(6), 1218–1252. <https://doi.org/10.1037/a0029334>

WCAG 2.0 conformance levels | UCOP. (n.d.). Retrieved April 13, 2022, from <https://www.ucop.edu/electronic-accessibility/standards-and-best-practices/levels-of-conformance-a-aa-aaa.html>

WCAG 2 Documents | Web Accessibility Initiative (WAI) | W3C. (n.d.). Retrieved March 1, 2022, from <https://www.w3.org/WAI/standards-guidelines/wcag/>

Winkler, R., & Soellner, M. (2018). Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. *Academy of Management Proceedings*, 2018(1), 15903. <https://doi.org/10.5465/ambpp.2018.15903abstract>

Zhou, Y., Zhou Gladys, Y. W., & Patton, D. H. (2019). A Mixed Methods Model of Scale Development and Validation Analysis. *Https://Doi-Org.Ezproxy2.Utwente.Nl/10.1080/15366367.2018.1479088*, 17(1), 38–47. <https://doi.org/10.1080/15366367.2018.1479088>

Appendix A

Focus Group Protocol

This appendix includes the protocol that was prepared and shared prior to the focus group formulation. This plan was devised collaboratively by the main researcher and Anna Boyko, an associate researcher who was involved in the first stage of this study. Below the plan can be seen.

Focus Group Protocol

Scheduled on Monday, Apr 4, 2022

Interviewer: Eric Kellenberger

Main Researchers: Maria Hristova, Anna Boyko

Duration: approximately 60 minutes

Preparation

Aim:

Test the two versions of the questionnaire and get the participants` opinion on the different design choices and identify their preferences in design.

Setting:

A room without possible distractions that can interrupt the flow of the conversation

Prior the focus group:

Distribute materials and informed consent form to the participants. The informed consent form will be provided in addition to this document.

Introduction and Tasks

Introduction (5 minutes):

Introduce the topic and the purpose of the study.

What can be expected to happen in the group?

Welcome, today we will discuss questionnaires assessing chatbots. First, we would like to show you a quick video about what a chatbot is.

<https://www.youtube.com/watch?v=pX6zqaEHAdw>


(show 45 sec of the video)

We will show you a chatbot and a possible example of a task you have to perform, after which you will be given two versions of the questionnaire. What we want from you is to fill in both of them and later on we want to discuss your view on which one you find better. It is also possible that you think both of them are good, both are flawed, or you find some elements in one good and others – better in the other version. We want to get your honest opinion on both versions. Also, we would like to stress that we are not interested in your opinion on the chatbot or in how you perform the task but in the understandability of the questionnaires you will be provided with.

Establish some ground rules:

- You can ask questions at any point if clarification is needed.
- If you encounter an issue with a question and prefer to state it immediately rather than later in the discussion, you can do so by informing the researchers and they can pay attention to your feedback. (Think aloud)
- We are only interested in your opinion about the questionnaire itself, not the chatbot.
- We will record the session as you have been informed already, but no information we will derive from this focus group will be shared with people outside of the research team and it will be anonymized.

Task (10 minutes):

Imagine you are in a Zoom meeting with a friend, preparing your homework together. You start experimenting with the settings in Zoom and remember seeing different backgrounds on other people when they use Zoom. Unfortunately, your friend also does not know how to change the background. So, you decide to get help on this via the Zoom website. Now, your task is to find the chatbot and ask for help. Please open the official Zoom website and look for the chatbot function. Often it is a chat symbol popping up in the corner like the one you can see here . Then, try finding needed information through the suggestions that the chatbot provides you with. After finding the video with the instruction, your task is finished.

Good luck!

Filling in questionnaires with think aloud (20 minutes):

After finishing the task, we ask you to fill in both questionnaires laying on your desk. You can think aloud and mention your opinion on the understandability and design of the questionnaire while filling it in.

Discussion (20 minutes):

Specific information we need/ Probes:

- Is the flow of the questionnaire easy to follow?
- Are there particular questions that are hard to answer?
- What are some design choices you like in the questionnaire? (Layout, answering options, even font and size?)
- What are some design choices you would like to be different?
- If you can state a clear preference for one questionnaire over the other, can you do so?
 - If not, point out two or three different things you like in each one?

Wrap up (5 minutes):

Thank you for your participation, do you have any further comments or questions?

Appendix B

Consent Form for Participation in the Focus Group

This appendix contains the consent form that was distributed to participants and if needed to their legal guardians to obtain their informed consent for participation in the study.

Informed consent form template for research with human participants

Authors: BMS Ethics Committee with input from Human Research Ethics TU Delft

Last edited: 20-01-2022

This consent form is associated with your participation in a focus group regarding the design of questionnaires for providing feedback on chatbots. You will be shown a short introductory video of what a chatbot is, after which you will be given a task to perform. Two versions of a scale for perceived usability and user satisfaction in chatbots, for shorter called the BotScale, will be given to you.

The aim of the focus group is to gather information about the design of the scale and its perceived accessibility. Your performance on the task and your opinion of the chatbot's usability will not be the main interest of the researchers.

Consent Form for Redesign for Accessibility of the Perceived Usability and User Satisfaction in Chatbots BotScale

Please tick the appropriate boxes

Yes No

Taking part in the study

I have read and understood the study information dated [*include date once it is confirmed*], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

I understand that taking part in the focus group involves my answers being audio recorded.

Note: The audio recording will be partially transcribed, and all names and identifiers of participants will be removed before usage. When the research purposes have been fulfilled, the audio and the transcript will be disposed.

Use of the information in the study

I understand that information I provide will be used by the research team to identify possible design flaws in the questionnaire.

I understand that personal information collected about me that can identify me, such as [e.g. my name or where I live], will not be shared beyond the research team.

Consent to be Audio Recorded

I agree to be audio recorded.

Signatures

_____	_____	_____
Name of participant	Signature	Date

and legal representative If applicable)

For participants unable to sign their name, mark the box instead of sign

I have witnessed the accurate reading of the consent form with the potential participant and the individual has had the opportunity to ask questions. I confirm that the individual has given consent freely.

_____	_____	_____
Name of witness	Signature	Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____	_____	_____
Researcher name [printed]	Signature	Date

_____	_____	_____
Researcher name [printed]	Signature	Date

Study contact details for further information:

Maria Hristova

m.hristova@student.utwente.nl

Anna Boyko

a.boyko@student.utwente.nl

Contact Information for Questions about Your Rights as a Research

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than


the researcher(s), please contact the Secretary of the Ethics Committee/domain Humanities & Social Sciences of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-hss@utwente.nl

Appendix C

Task Formulation for Zoom Chatbot

This is the task that was used in both the focus group and the quantitative data collection study design.

Task Instructions

Imagine you are in a Zoom meeting with a friend, preparing for your work together. You start experimenting with the settings in Zoom and remember seeing different backgrounds on other people when they use Zoom. Unfortunately, your friend also does not know how to change the background. So, you decide to get help on this via the Zoom website. Now, your task is to find the chatbot and ask for help. Please open the official Zoom website and look for the chatbot function. Often it is a chat symbol popping up in the corner like the one you can see here . Then, try finding needed information through the suggestions that the chatbot provides you with. After finding the video with the instruction, your task is finished, and you can come back here.

Please follow the link to conduct the task: <https://zoom.us/>

Remember that we are interested in your experience with the chatbot, if for any reason you cannot achieve the goal in a reasonable amount of time, please simply come back here once that you gain enough knowledge to assess the quality of the chatbot.

Appendix D

This appendix contains the design referred to as D1 in the focus group.

Demographics

Before actually conducting the questionnaire, you are asked to fill in some questions about yourself.

1. How old are you?

Please fill in a number.

2. What is your disability?

Please fill it in.

3. Are you already familiar with chatbots?


Please tick the right answer.

Yes

No

Chatbot Usability Questionnaire

Instructions

In what follows, you will first be asked to fill in some information about yourself. Then, you will find 11 statements regarding the chatbot you have just used.  Please provide your honest opinion on how strong you agree with the statements. The answer possibilities range from “strongly disagree” to “strongly agree”.

To answer, fill in the button you agree with the most. Here, you can find an example questions with an example answer:

1. This sentence is written in English.

1



2



3



4



5



STRONGLY
DISAGREE

DISAGREE

NEITHER
DISAGREE NOR
AGREE

AGREE

STRONGLY
AGREE

2. It was easy to find the chatbot. 🔍

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE NOR
AGREE

4



AGREE

5



STRONGLY
AGREE

3. Communicating with the chatbot was clear. ✓

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE NOR
AGREE

4



AGREE

5



STRONGLY
AGREE

4. The chatbot was able to keep track of context. ✓

1 

STRONGLY
DISAGREE

2 

DISAGREE

3 

NEITHER
DISAGREE NOR
AGREE

4 

AGREE

5 

STRONGLY
AGREE

5. The chatbot's responses were easy to understand. ✓

1 

STRONGLY
DISAGREE

2 

DISAGREE

3 

NEITHER
DISAGREE NOR
AGREE

4 

AGREE

5 

STRONGLY
AGREE


6. I find that the chatbot understands what I want and helps me achieve my goal. 

1 


STRONGLY
DISAGREE

2 


DISAGREE

3 

NEITHER
DISAGREE NOR
AGREE

4 

AGREE

5 

STRONGLY
AGREE


7. The chatbot gives me the appropriate amount of information. 

1 


STRONGLY
DISAGREE

2 


DISAGREE

3 

NEITHER
DISAGREE NOR
AGREE






4 

AGREE






5 

STRONGLY
AGREE

8. The chatbot only gives me the information I need. 

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1 	2 	3 	4 	5 
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

9. I feel like the chatbot's responses were accurate. 

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1 	2 	3 	4 	5 
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

10. I believe the chatbot informs me of any possible privacy issues. 

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE NOR
AGREE

4



AGREE

5



STRONGLY
AGREE

11. My waiting time for a response from the chatbot was short. 

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE NOR
AGREE

4



AGREE

5



STRONGLY
AGREE

You have finished the questionnaire. Thank you for your participation!

Appendix E

This appendix contains the design referred to as D2 in the focus group.

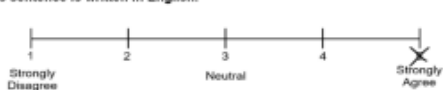
Chatbot Usability Questionnaire

Section one instructions:

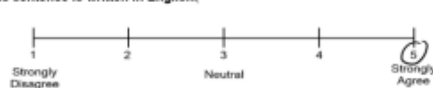
In this section you are given three sentences. Choose only one of the possible answers on how much you disagree or agree with the sentence. Put **X** or **O** over the number of your choice.

This is an exemplary sentence with an answer:

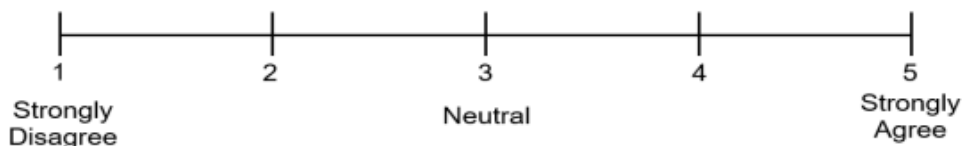
This sentence is written in English.



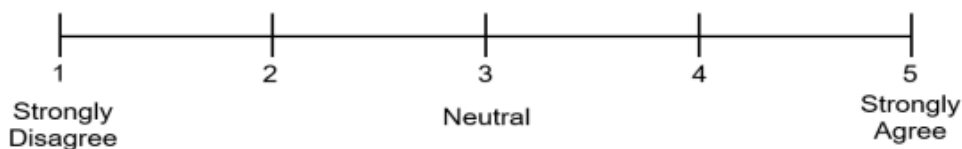
This sentence is written in English.



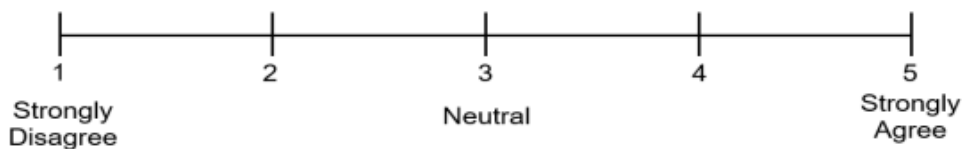
1. I am familiar with chatbots or other conversational agents.



2. I know how chatbots work.



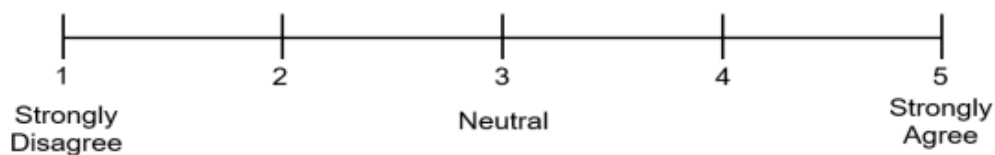
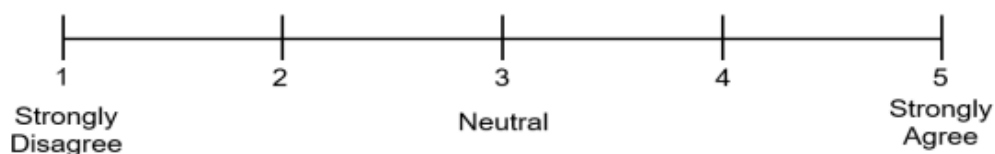
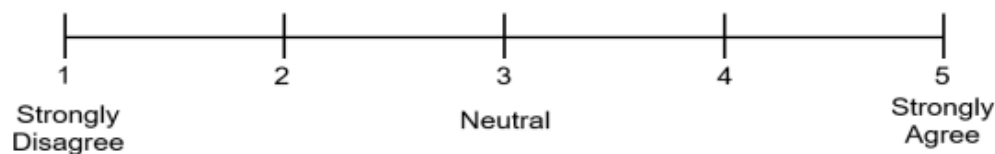
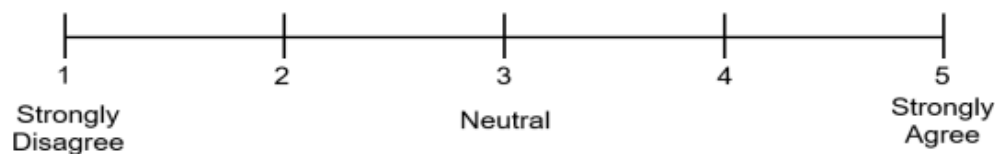
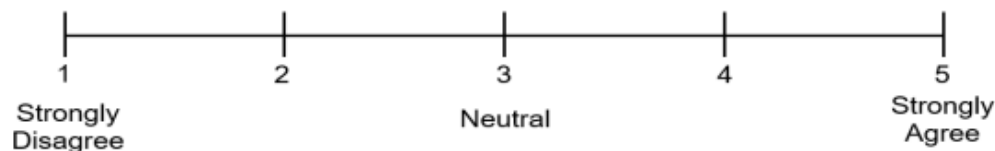
3. I am confident in using chatbots.



If you have performed the task provided to you fill in the remaining part of the questionnaire. If not, please perform the task and then come back to the questionnaire.

Section two instructions:

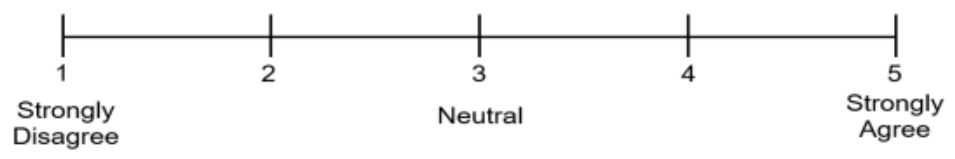
You will be given 11 statements to read. For each statement choose only one of the possible answers on how much you disagree or agree with the sentence. Put **X** or **O** over the number of your choice.

1. The chatbot function was easily detectable.**2. It was easy to find the chatbot.****3. Communicating with the chatbot was clear.****4. The chatbot was able to keep track of context.****5. The chatbot's responses were easy to understand.**

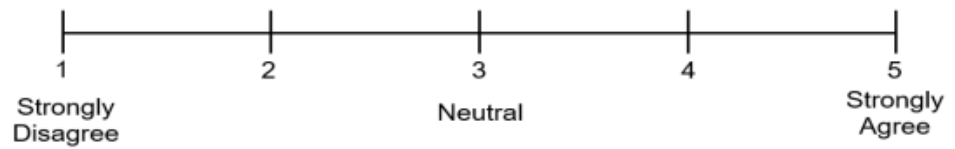
Continue to next page

Continuation of previous page

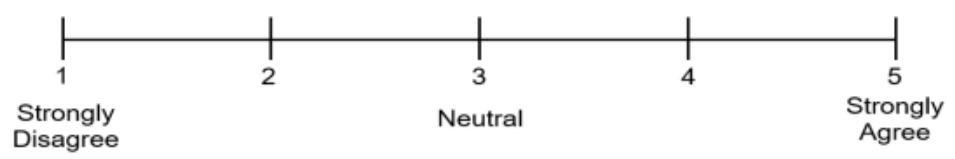
6. I find that the chatbot understands what I want and helps me achieve my goal.



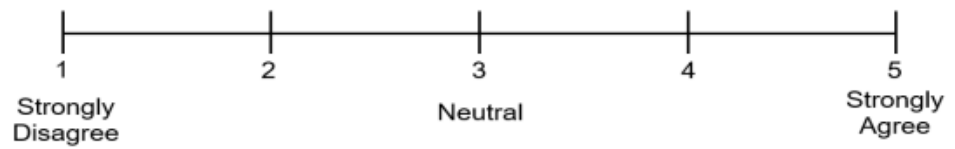
7. The chatbot gives me the appropriate amount of information.



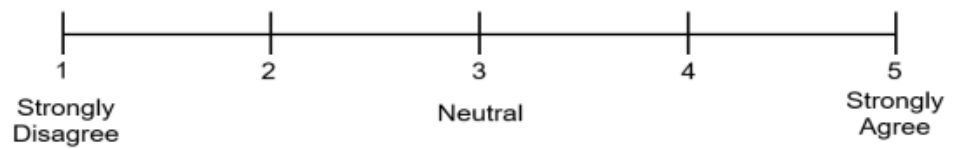
8. The chatbot gives me the information I need.



9. I feel like the chatbot's response was accurate.



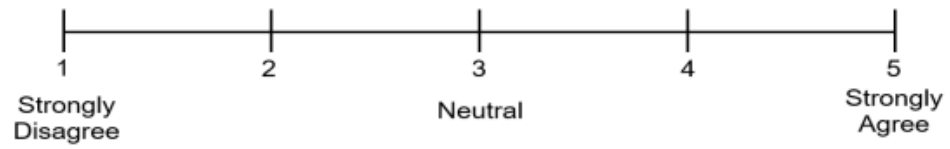
10. I believe the chatbot informs me of any possible privacy issues.



Continue to next page

Continuation of previous page

11. My waiting time for a response from the chatbot was short.



This is the end of the questionnaire! Thank you for your participation.

Appendix F

Focus Group Transcript

This is the transcript from the focus group. All participants and their communication partners, as well as the associated researchers have been granted their anonymity.

Open Mind School
 Chatbot Accessibility Focus Group
 Interview 1
 Interviewer: Associate researcher 1
 Transcriber: Associate researcher 1

Interviewer	I
Communication Partner	C
Participant 1	P1
Participant 2	P2

P1 [16:35 in recording]

I: Was the flow of the questionnaire easy to follow?

C: Do you think it was easy or hard to follow?

P1: Easy

I: Okay, easy. Are there particular questions that were hard to answer?

C: Did you find that there were questions that were harder, yes or no?

P1: No.

I: No? Okay. What are some design choices you like in the questionnaire? Did you like the layout?

P1: Yes.

I: Did you like the answering options?

P1: No.

I: No, okay. Did you like the font and size?

P1: No.

I: Can you tell me, what would you change about the font and size?

C: Would you want it bigger or smaller?

P1: Bigger.

I: Okay, bigger. What are some design choices you would like to be different? You said you didn't like the answering options. What did you not like about them? Did you not like how they were worded?

C: Did you like the way they were worded?

P1: Yes.

I: Would you change the scale?

P1: No.

I: How was the length of the questionnaire? Was it too short, about right, or too long?

P1: Too long.

I: Too long, okay. Good to know. And you think both questionnaires were too long?

P1: No.

C: The first one or the second one was too long?

P1: First one. [Grayscale]

I: Okay, thank you. Did you have a preference for one over the other?

C: Did you like the first one or second one more?

P1: Second one. [Color]

P2 [21:44 in recording]

I: Was the flow of the questionnaire easy to follow?

P2: Yes.

I: Yes? Were there any particular questions that were hard to answer?

P2: Yes.

I: Which questions were hard to answer? Can you show me?

P2: That one and this one.

I: Were there any more?

P2: No.

I: Just the first two?

P2: Yeah.

I: Okay. So "I am familiar with chatbots" and "I know how chatbots work" were hard to answer. What were some design choices that you liked about the questionnaire?

Everything.

I: Did you like the layout?

P2: Yes.

I: Between the first one and the second one, which one did you like better?

P2: Everything.

I: Did you say everything?

P2: Yeah.

I: So you didn't have a preference between 1 and 2?

P2: No.

I: Okay. How was the font and size?

P2: Bad.

I: Bad? How would you change it?

P2: I don't know.

I: Would you make the text smaller or bigger?

P2: Bigger.

I: Bigger, okay. What about the answering options; did you like it going from 1 to 5?

P2: Yes.

I: Are there other design choices that you would want to be different?

P2: Yes.

I: Can you tell me about those?

P2: Yes.

I: For example, did you like or dislike having the images of the smiley faces?

P2: Yes.

I: Did you like or dislike it?

P2: Liked it.

I: You liked it, okay. Did you like or dislike that there was color?

P2: Liked.

I: You liked that there was color. Can you state if you have one of the questionnaires that you liked better than the other? Did you like the first questionnaire better or the second questionnaire better?

P2: Second. [Color]

I: Can you tell me, what is the main reason you like the second one better?

P2: Because I like it.

I: Because you like it. Well you said you like the color and you said you like the smiley faces. Was there anything else you liked about it?

P2: No.

Appendix G

This appendix contains the finalized design of the BUS-A scale.

BUS-A

BOT USABILITY SCALE ACCESSIBLE VERSION



CREATED BY

ANNA BOYKO a.boyko@student.utwente.nl

MARIA HRISTOVA m.hristova@student.uwente.nl

IN COLLABORATION WITH

DR. SIMONE BORSCI s.borsci@utwente.nl


ERIC KELLENBERGER eric@openmindschool.org

This questionnaire is assigned to participant number . This number is to be filled in on the top right corner of every page.


N

Chatbot Usability Questionnaire

Instructions

In what follows, you will first be asked to fill in some information about yourself. Then, you will find 12 statements regarding the chatbot you have just used. 

Please provide your honest opinion on how strong you agree with the statements. The answer possibilities range from “strongly disagree” to “strongly agree”.

To answer, fill in the button you agree with the most. 

Demographics

Before actually conducting the questionnaire, you are asked to fill in some questions about yourself.

1. How old are you?

Please fill in a number (e.g. 18 if you are eighteen years old).

2. Do you consider yourself to have a disability that can affect your experience with the chatbots (e.g. vision problems that are not corrected with glasses, learning disability, mental health or emotional disability, unseen disability, physical disability, sensory disability, etc.)?

Please tick the right answer.

Yes

No

3. If yes, how would you describe your disability? Please tick as many as apply to you.

*Information associated with this question is not going to be used or shared for the research

**This question is optional and could be skipped

- Developmental Disability
- Learning disability
- Mental health or emotional disability
- Unseen disability
- Physical disability
- Sensory disability
- If you use an alternative term, please describe here:

Decline to answer

4. What is your current gender identity? (check all that apply)

*Information associated with this question is not going to be used or shared for the research

** This question was developed in tune with: Broussard, K. A., Warner, R. H., & Pope, A. R. (2018). Too many boxes, or not enough? Preferences for how we ask about gender in cisgender, LGB, and gender-diverse samples. *Sex Roles*, 78(9), 606-624.

Man

Female

Female-To-Male (FtM) / Transgender male / Trans male

Male-To-Female (MtF) / Transgender female / Trans woman

Genderqueer, neither exclusively female nor male

Additional Gender Category (Other), please specify:

Decline to answer

5. What was your sex as assigned at birth?

Please tick the right answer.

Male

Female

6. Are you already familiar with chatbots?

Please tick the right answer.

Yes

No

You have finished the demographic questionnaire. Now, the questionnaire about the chatbots will be presented.

Bot Usability Scale

In what follows, you will be asked to rate your agreement to the 11 statements from 1

“strongly disagree” to 5 “strongly agree”. To answer, fill in the button you agree with the most.

Please choose **only one** answer.

One statement will be presented on one page.

Here, you can find an example question with an example answer:

1. This sentence is written in English.

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE
NOR AGREE

4



AGREE

5



STRONGLY
AGREE

1. The chatbot function was easily detectable. 🔍



1



STRONGLY
DISAGREE



2



DISAGREE



3



NEITHER
DISAGREE
NOR AGREE



4



AGREE



5



STRONGLY
AGREE

2. It was easy to find the chatbot. 🔍



1



STRONGLY
DISAGREE



2



DISAGREE



3



NEITHER
DISAGREE
NOR AGREE



4



AGREE








5








STRONGLY
AGREE






3. Communicating with the chatbot was clear. ✓

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5
				
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

4. The chatbot was able to keep track of context. ✓

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5
				
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

5. The chatbot's responses were easy to understand. ✓

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5
				
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

6. I find that the chatbot understands what I want and helps me achieve my goal.



1

2

3

4

5



STRONGLY
DISAGREE

DISAGREE

NEITHER
DISAGREE
NOR AGREE

AGREE

STRONGLY
AGREE

7. The chatbot gives me the appropriate amount of information. 🎯

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE
NOR AGREE

4



AGREE

5



STRONGLY
AGREE

8. The chatbot only gives me the information I need. 🎯



9. I feel like the chatbot's responses were accurate. 🎯



1



STRONGLY
DISAGREE



2



DISAGREE



3



NEITHER
DISAGREE
NOR AGREE



4



AGREE







5



STRONGLY
AGREE

10. I believe the chatbot informs me of any possible privacy issues. 

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5
				
STRONGLY DISAGREE	DISAGREE	NEITHER DISAGREE NOR AGREE	AGREE	STRONGLY AGREE

11. My waiting time for a response from the chatbot was short. 🕒

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE
NOR AGREE

4



AGREE

5



STRONGLY
AGREE

You have finished the questionnaire. Thank you for your participation!

Appendix H

Consent Form

This was the consent form used in the second phase of the study.



Introduction to the study

Dear Participant,

You are invited to take part in study conducted by Anna Boyko and Maria Hristova, in collaboration with Mustafa Taha, supervised by Dr. Simone Borsci. The overall aim of this part of the study is to assess your satisfaction using chatbots to find information. We will use your answer to validate the scales for the assessment of chatbot quality.

Overall, the study will last approximately 20 minutes and data will be anonymized to protect your identity.

Requirements for participation

To be eligible for the control group of this study, it is important to have proficient **English skills**.

Risks associated with the study

There are no known risks associated to the study.

Benefits associated with the study

There are no personal benefits associated with participating in the study. Nevertheless, we conduct the study aiming to make research more inclusive. By conducting the study, you could help coming one step closer towards that aim.

Confidentiality

Your answers will be completely anonymous. We will not be able to identify you based on your answers. In the data analysis, your answers will be treated anonymously too. Your answers will only be used for the purpose of the study.

Disclaimer

Your participation is voluntary, and you can withdraw from the study at any time.

P.S.: This survey contains a completion code for SurveySwap.io

I hereby declare I meet the requirements for participation.

- Yes, I meet the requirements.
- No, I do not meet the requirements.