

Design and Evaluation of an Inclusive Version of the Bot Usability Scale

Anna Boyko

s2284774

a.boyko@student.utwente.nl

Department of Cognitive Psychology and Ergonomics

University of Twente

First Supervisor: Dr. Simone Borsci

Second Supervisor: Jule Landwehr

June 30, 2022

Abstract

Background. People with disabilities face several issues in accessibility which get extrapolated by excluding them from usability research and hence from improving their situation. Therefore, satisfaction scales testing user experience need to be utilized to get insights into such points of improvement. A scale, currently under validation is the Chatbot Usability Scale (BUS-11). It aims at measuring satisfaction with chatbots, which are implemented more and more in customer service and for interventions. **Objective.** The current study had the objective to design an accessible version of the BUS-11 and to then test and validate it. **Method.** After testing two prototype versions of the accessible BUS-11 in a focus group, a final version was created (BUS-A). To test the psychometric quality of the new version of the scale, participants with and without disabilities were asked to interact with the same chatbot, and assess their satisfaction using one of the two scales (BUS-11 or BUS-A) in a between-subject study design. Participants with disabilities always used the BUS-A (Group 1), participants without disabilities were assigned randomly to BUS-A (Group 2) or BUS-11 (Group 3). T-tests, regression analyses, and confirmatory factorial analyses were implemented to test for differences between the scales, and differences in satisfaction scores between the three different groups. **Results.** T-test analysis showed a significant difference between the satisfaction scores of participants who used the BUS-11 and the BUS-A ($t(39.682) = -2.235$, $p = .031$). However, the linear regression showed no significant differences between the three groups ($F(2, 113) = 2.643$, $p = 0.076$, $R^2 = 0.028$). A t-test between people with and without disabilities conducting the BUS-A also showed no significant differences ($t(5.070) = -0.272$, $p = .796$). Lastly, CFA confirmed the factorial structure as anticipated (CFI = .941, RMSEA = .080, SRMSR = .048). **Conclusion.** Overall, the current study showed that the BUS-A can be considered accessible, although outcomes show that the scale might be treated as a completely different assessment instrument than the BUS-11.

Keywords: Disability, Accessibility, Chatbots, Usability, Chatbot Usability Scale (BUS)

Contents

Abstract	1
Contents	2
Introduction	4
Importance of Usability Testing for Chatbots and Barriers to Inclusive Research	7
Limitations of Different Disabilities and Principles of Inclusive Design	10
Aims of the Current Study	12
Phase 1 – Design an Accessible Version of the Chatbot Usability Scale (BUS-A) and Focus Group	13
Principles to Drive the Design	14
Focus Group About the BUS-A Prototypes	22
Participants	22
Materials	23
Procedure	23
Data Analysis	24
Results	24
Discussion Phase 1	26
Testing BUS -A and Comparative Analysis with BUS-11	27
Methods	27
Participants	27
Materials	28
Procedure	29
Data Analysis	29
Results	31
Descriptive Statistics	31
Hypothesis Testing	33
Psychometric Assessment of the Scales	34
Reliability Analysis of the Different Versions of the Scale	37
Discussion Phase 2	39
Overall Discussion	41
Future Research	45
Conclusion	47
References	48
Appendix A - Informed Consent Focus Group	56
Appendix B - Focus Group Protocol	59
Appendix C - Task Instructions Focus Group	62
Appendix D - Prototype BUS-A 1	63

Appendix E - Prototype BUS-A 2.....	67
Appendix F - Transcription of the Focus Group	71
Appendix G - Final Version – BUS-A	73
Appendix H - RStudio Code	83

Introduction

People with disabilities encounter several barriers in their daily life. Not only do they suffer social exclusion in general by facing accessibility issues to participate in all sorts of activities in social communities (Kissow, 2013), but they also lack the possibilities to work on improving this situation of social exclusion (Abbott & Mcconkey, 2006). Abbott and Mcconkey (2006) point out that it is a crucial step toward inclusivity to include opinions and voices of the disabled, and thus to work together with them. However, disabled people often stay dependent on others to identify factors for them on what to do about it (Abbott & Mcconkey, 2006). Partly, this could be due to their lacking possibilities to participate in research. Vereenoghe (2021) reported that in a study of 300 participants with disabilities, 90% got excluded from research trials, although 70% could have participated if accommodations in the study design would have been made. Thus, people with disabilities often do not get the chance to conduct questionnaires as these are designed without any special accommodations for their needs (Schrepp et al., 2016). If they however would have the possibility to participate in research, they could also participate in usability research, which could then be used for extrapolating insights into their experiences and needs.

To get insights into difficulties and needs people with disabilities might have on a daily basis, a possible path towards social inclusion could be usability research. When it comes to the interaction with digital systems, usability is defined as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” under ISO 9241-11 standard (ISO, n.d.). In line with this standard definition of usability, satisfaction is the subjective metric of the “extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user’s needs and expectations“ (ISO, n.d.). Effectiveness is defined as “accuracy and completeness with which users achieve specified goals” and efficiency as “resources used in relation to the results

achieved” (ISO, n.d.). By testing usability and its related concepts, important points of improvement toward social inclusion could be revealed and adapted accordingly in the design.

A software application which is currently not only gaining increasing interest from companies in different fields (Valério et al., 2017) but also shows great potential for helping people with disabilities, are chatbots (Arias-Durán et al., 2021). Chatbots, aimed at interacting live with users, are supposed to increase customer productivity on the one side and reduce service costs on the other side (Valério et al., 2017). However, even though representing such great technological development and economic benefits, accessibility does not develop in tune with functionalities (Torres et al., 2018). This is not only an example of a social service where inclusivity is lacking but also an example of a potentially beneficial system that could be used to help people with disabilities in various ways (Arias-Durán et al., 2021). For instance, chatbots could become of great importance by providing real-time interventions in disabled people’s natural environments. They could provide immediate advice and support for people in need (Arias-Durán et al., 2021). Making chatbots accessible would thus not only open up a daily service for people with disabilities but could also improve their lives (Arias-Durán et al., 2021). Therefore, next to accessibility, usability is crucial to improve the ease and quality of use of those and thus attract users (Arias-Durán et al., 2021). However, little is known so far on how to assess the levels of effectiveness and efficiency of chatbots (Valério et al., 2017), let alone on how to make those inclusive for everyone.

A means to measure parts of usability, including the usability of chatbots, are satisfaction scales. Probably the most popular scales for assessing satisfaction by taking a closer look at perceived usability are the System Usability Scale (SUS) (Lewis, 2018b), the Computer System Usability Questionnaire (CSUQ) and the Usability Metric for User Experience (UMUX) (Lewis, 2018a). Furthermore, a scale correlating with the shorter version of the UMUX (UMUX-Lite), currently under validation, is the Chatbot Usability Scale (BUS-

11) (Borsci et al., 2021b). Unlike the presented general scales, the BUS-11 measures usability for chatbots specifically, which could aid in more specific insights of improvement (Borsci et al., 2021b).

Borsci et al. (2021b) created this new tool, the BUS-11, to test the quality of interaction with chatbots in a standardized manner. As this scale aims at assessing satisfaction after the interaction with chatbots, thus, providing feedback to the designers on how to improve the chatbot, it is important to enable people with disabilities to use the BUS-11 so that the needs and the perspectives of people with different levels of individual functioning are going to be considered in designing and improving chatbots.

The validated version of the BUS is composed of 11 items measured on a 5-point Likert scale, divided into 5 factors with an overall level of reliability equal to 0.9. As can be seen in Table 1, the factors aim at investigating perceived accessibility to the chatbot functions (Factor 1), their perceived quality overall (Factor 2), but also the perceived quality of the conversation and the chatbot provided (Factor 3). In addition, factor four investigates perceived privacy and security and factor five is the time response. Thereby, aiming to examine a complete picture of the interaction with the chatbot.

Table 1

Factors and Items of the Bot Usability Scale

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 functions	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	<p>6. I find that the chatbot understands what I want and helps me achieve my goal.</p> <p>7. The chatbot gives me the appropriate amount of information.</p> <p>8. The chatbot only gives me the information I need.</p> <p>9. I feel like the chatbot's responses were accurate.</p>
4 - Perceived privacy and security	10. I believe the chatbot informs me of any possible privacy issues.
5 - Time response	11. My waiting time for a response from the chatbot was short.

The current version of BUS-11 was designed as a classic Likert scale and therefore it only partially fulfils the needs of people with disabilities. This is due to the special requirements regarding literacy and design needed for disabled people to successfully conduct questionnaires (Davies et al., 2017). To make the scale more inclusive, accessible adaptations and procedures should be followed during the design (Goegan, 2018; Fuchs et al., 2005). These adaptations needed are not in line with the current version of the BUS-11. Considering that questionnaires for the evaluation of chatbots are rare (Borsci et al., 2021b), but those could have a positive impact on the support of disabled people, the current study aims at designing a version of the BUS-11 scale accessible for people with disabilities.

Importance of Usability Testing for Chatbots and Barriers to Inclusive Research

Chatbots, which are used in a variety of different contexts, such as in business, education and healthcare have gained increasing interest during the last years (Zumstein & Hundertmark, 2017). Generally, they can be described as conversational software agents

(CAs) that employ natural language processing and can simulate human language to respond to real individuals (Adam et al., 2020). The artificial intelligence of the software enables 24/7 personalized support for people all over the globe (Sanny et al., 2020).

Despite being implemented in those different contexts, one especially profitable application of chatbots seems to be customer support in various industries (Adam et al., 2020). Here, chatbots are expected to save more than \$8 billion per year by 2022 (Adam et al., 2020). These savings are a result of more efficient and productive solutions than traditional customer service can provide (Sanny et al., 2020). Chatbots are available 24/7, respond directly and resolve the need to scroll through pages of questions and answers (Adamopoulou & Moussiades, 2020). Thereby, aiming to reduce time and effort in looking for answers.

Although chatbots are being presented with such high potential in customer service, these benefits are not inclusive to everyone. Even though the launch of the web in the 1990s led people to quickly recognize that adjustments to include users with disabilities need to be made (Rømen & Svanæs, 2008), people with disabilities still face a variety of usability issues on the internet (Spina, 2019). Moreno et al. (2012) report that especially problems with reading and understanding hinder disabled people from efficient internet use. In this context, attention should not only be paid to usability, but also to accessibility, which aims at “expanding the range of users who can successfully access a website” (Moreno et al., 2012). As Schrepp et al. (2016) suggest, all relevant user groups should be able to have access to the full user experience. As chatbots are often designed without accessibility standards in mind from the beginning (Torres et al., 2018), the reality is still far from this.

Not only does the inaccessibility of chatbots hinder disabled people from experiencing the same presented benefits as people without disabilities, but it can also be seen as a wasted opportunity. This is not exclusive to chatbots but is often the case for a variety of technologies. Where potential is shown to help disabled people to overcome their barriers and

disparity in everyday life, some technologies fail to fulfil this potential due to a lack of usability for everyone (Tsatsou, 2020). Moreover, reality often shows rather unsatisfactory results when engaging with certain chatbots in general. Thus, they do not only exclude people with disabilities from efficient use, but they also show problems in other facets of usability. Adam et al. (2020) even put the question forward whether these chatbots will be effective in the long run. Thus, testing is needed to identify points of improvement (Silderhuis, 2020). Still, standardized scales to measure users' satisfaction are rare (Balaji, 2019).

Testing metrics relating to usability is crucial to get insights into the performance of a system or application (Silderhuis, 2020). Moreover, customer satisfaction as a factor has been reported to be closely connected with the performance of a company in general (de Haan et al., 2015) and also have been shown to be superior to other user experience-related measurements (Kvale et al., 2021). If testing usability metrics, the quality of chatbots can be increased and people can use them more efficiently (Arias-Durán et al., 2021). Furthermore, especially inclusive research in collaboration with disabled people has proven to result in impactful outcomes (Walmsley, 2001). Including disabled people in usability assessments could increase the range of users able to operate chatbots. Consequently, this would also enable research with more diverse samples, as disabled participants could also contribute to usability research if scales would be made accessible to them. As indicated by Saenz et al. (2017), one way to measure user satisfaction with chatbots in a standardized, quantitative manner is using usability scales designed for web and technologies in general.

As already referred to, some of the most widely used scales appear to be the SUS, the CSUQ and the UMUX (Lewis, 2018a; Lewis, 2018b). Even though all scales show good psychometric properties (Lewis, 2018a), the SUS has been shown to be the most popular to be used (Lewis, 2018b). The shorter version of the UMUX scale, the UMUX-Lite, containing two items, is considered as having especially high reliability estimates ($\alpha = .82$) (Lewis et al., 2013). Even though these scales show high reliability in measuring usability and customer

satisfaction in general, they are neither specified in measuring customer satisfaction with chatbots nor are they inclusive to everyone.

Given that no standardized questionnaires were available to test chatbots specifically, Borsci et al. (2021b) created the BUS-11 scale specifically tailored to measure their usability. The confirmatory analysis (Borsci et al., 2021b) showed high reliability of the 11-item scale ($\alpha = .89$). Further studies have already evaluated the scale's reliability, resulting in satisfactory coefficients (Kerwien Lopez, 2021; Bos, 2021). Also, the BUS-11 was shown to be highly correlated with the UMUX-Lite ($r_t = .68, p < .001$), showing high potential (Borsci et al., 2021b). Still, the usability scale is under investigation, requiring further rounds of testing and validation (Borsci et al., 2021a). Also, special design changes need to be implemented to enable disabled people to participate in the evaluation of chatbots and thus increase the range of users able to operate chatbots.

Limitations of Different Disabilities and Principles of Inclusive Design

Considering that approximately 1 billion people, translating into 15% of the world population, are living with some sort of disability (World Health Organization, n.d.), their strengths and weaknesses need to be taken into account when designing for them. Harper and Chen (2011) found that guidelines for more inclusive design are often ignored. Moreover, guidelines for more accessible and inclusive design were followed only in 10% of the cases over 10 years (Harper & Chen, 2011). Different impairments require specific sets of adaptations, which can increase usability for disabled people (Vanderheiden et al., 2021). Keeping in mind the goal of making surveys and research more inclusive, certain guidelines can be identified to achieve this goal, including those which do not specifically aim at surveys.

Generally, there is no approach to designing that results in a fit for everyone. By considering people's weaknesses in the design stage of constructing a questionnaire or testing instrument, usability can be increased. For example, visual impairments can be considered in

the design by incorporating larger letters and high colour contrasts (Vanderheiden et al., 2021). Thereby, improving overall readability (Vanderheiden et al., 2021). An important target group to consider when designing questionnaires are people with learning disabilities (Vanderheiden et al., 2021). These neurodevelopmental disorders are summarized by DSM-5 and typically include ongoing difficulties in reading, writing and/or maths (American Psychiatric Association, n.d.). Moreover, children with ADHD and autism, portraying problems with attention, executive functioning and processing speed, also show learning differences (Mayes & Calhoun, 2007). Considering their main difficulties, these groups of people are especially important in including in usability design.

In an attempt to make such guidelines standardized and clear, the World Wide Web Consortium (W3C) provides instructions to make the web more accessible (Lewthwaite, 2014). Here, the four main principles to be followed are (1) Perceivable, (2) Operable, (3) Understandable and (4) Robust (W3C, 2018). The first point, perceivable, includes that all the information and its components need to be presented in a way that everyone can understand. This can be achieved through, for example, text alternatives. Operable means that all aspects need to be functional by, for example, providing enough time to use them. Third, understandable includes design recommendations ensuring that everyone recognizes the information given and understands the different components. Here, design principles such as bigger letters and contrasts, which were already discussed, are incorporated. Lastly, the W3C states that design needs to be robust, meaning that content needs to be interpreted by different user agents, such as assistive technologies the same way (W3C, 2018). All in all, some of the more specific recommendations of these four main principles can also be applied in making research more accessible.

Another attempt to increase the accessibility of research for people with disabilities includes incorporating accommodations in setting, timing, presentation and response (Goegan, 2018). By for example controlling the environment, such as allowing for breaks in

between, presenting the data through different modes and altering the Likert scale to, for example, seven points, certain barriers could already be removed for people with different disabilities (Goegan, 2018). An example of such modified questionnaires is presented by the company easy-read-online. They describe using plain text, pictures, a clear layout and smileys on the scale to develop their questionnaires (Easy Read Online, n.d.). As can be seen, there are already attempts and guidelines made for the inclusive design of the web and questionnaires in research. What remains an important aim of the current study is to remember that one should design in collaboration with people with specific disabilities, when designing for them.

Aims of the Current Study

The aim of the current study is threefold: i) design an accessible version of the Bot Usability Scale (BUS-A) in line with the guidelines and principles of inclusive design (Goegan, 2018; Vanderheiden et al., 2021; W3C, 2018) ii) evaluate how potential users with disabilities perceive their experience using the redesigned tool in the second step as a basis for the final design, iii) compare the psychometric properties of the final version, the BUS-A, with the original scale and its factorial structure, analysing differences between the two scales.

Currently, the original scale is not inclusive of different target groups with special needs. Therefore, the first step to achieving the goals will be to implement and test a new (accessible version) of the BUS-11 that is called BUS-A. As it is important to work in collaboration with disabled people (Walmsley, 2001), the prototype versions of the BUS-A will be tested partly in collaboration with the Open Mind School in California. This will be done to ensure the inclusion of participants with various disabilities and backgrounds in a reliable and controlled manner. The user experience of filling in the scale will be assessed qualitatively in a focus group with potential end-users of the scale, which serves as a basis for the final design decisions.

After the redesign of a more accessible version of the scale, this new scale will be used to collect data regarding the satisfaction with chatbots to evaluate the psychometric properties of the BUS-A and compare these properties with the original version of the scale. Thereby, special focus will be placed on comparing its reliability and factorial structure with the initial BUS-11 psychometric properties.

The following three research questions summarise the goal of the psychometric investigation comparing BUS-A and the BUS 11:

RQ1: Are the two scales, BUS-A and BUS-11 significantly different from each other in terms of participants' satisfaction rate with a chatbot?

RQ2: Do people with and without disabilities report significantly different satisfaction scores using the BUS-A?

RQ3: Are the psychometric properties of the BUS-A (factorial structure and reliability) in line with the one of the original scale?

Based on the previous literature review, expectations regarding the research questions can be formulated. As following pre-specified design principles should lead to an accessible design, imitating the structure of the original BUS-11, the corresponding expectations are that the two scales should not be significantly different from each other in terms of participants' satisfaction rate with a chatbot (RQ1), but also that people with and without disabilities do not report significantly different satisfaction scores using the BUS-A (RQ2). Furthermore, in line with the third research question, no differences in psychometric properties between the two scales are expected.

Phase 1 – Design an Accessible Version of the Chatbot Usability Scale (BUS-A) and Focus Group

The first phase of the study concerns designing an accessible version of the BUS-11. Two researchers designed two different prototypes of the BUS-A and explored these versions of the scale in a focus group with people with disabilities to make decisions regarding the

final design of the BUS-A. First, both researchers conducted research on principles separately from each other before merging all of them. Afterwards, the two prototype versions of the BUS-A were designed in collaboration. Independently from each other, one version resulted in a more minimalist design, without visual distractions (prototype 1), and the other one (prototype 2) included more colour and pictorials by following design recommendations different from the first prototype. This allowed researchers to get more qualitative insights into what the target group favoured more. The prototypes were designed in Word (Microsoft, n.d. a) and both the designs were developed in line with design recommendations. In accordance with the aim of redesigning the scale, Microsoft (n.d. a) suggests several steps and instructions on how to create accessible documents to achieve this goal.

Principles to Drive the Design

The prototypes of the BUS-A were designed by following several guidelines found in literature and online resources regarding accessible questionnaires, and accessible web and PDF design. Table 2 provides a detailed overview of the principles followed by each of the prototypes. Thereby, principles were divided into five categories to create a clearer overview: Media; Text; Navigation and Links; Colours; and Assistive Technologies (Spire Digital, 2019). Even though principles were derived from various sources of guidelines, they mostly overlap and result in a structured overview of points to follow in designing for accessibility (Table 2).

Principles defined under Media are a) to provide visual response alternatives as an addition to the typical Likert Scale response options (Hartley & MacLean, 2006), b) to make sure that alternate text is added to all images included in the design (University of Washington, n.d.; W3C, 2018) and c) to minimize elements that can cause visual distractions (Spire Digital, 2019). These principles derived from different guidelines ensure that the media conveying the questionnaire is perceivable to the end-user (W3C, 2018).

Almost all Guidelines included similar principles regarding the category Text: principles a), c) and e) were specified by guidelines of Spire Digital (2019), while b) and d) can be found in several sources of guidelines (Goegan et al., 2018; Vanderheiden et al., 2021). Again, following these guidelines ensures perceivability for any target group of users (W3C, 2018).

The third category of principles, Navigation and Links (Table 2) summarizes principles aiming at making the questionnaire operable (W3C, 2018). Thereby, each principle was derived from a separate guideline, even though they were mostly present in all guidelines. Principle 3a) asks for additional explanations to make sure that the user understands the navigation of the questionnaire and can operate it (Spire Digital, 2019). Principle b) suggests making use of Gestalt principles and grouping techniques in general to provide cues to which parts of the questionnaire belong together (Vanderheiden et al., 2021). Principle c) recommends using radio buttons instead of the typical table usually used for Likert scales on PDFs (Si, 2020), d) to convey information through multiple mediums if possible (Hartley & MacLean, 2006), and principle e) states to use Word's built-in headings so assistive technologies are also able to differentiate between headings and text.

Accessible implementation of colours is important for a perceivable design by making elements in the questionnaire distinguishable (W3C, 2018). Thereby, both principles under the fourth category relate to including accessible and contrasting colours (Microsoft, n.d. a). As contrast alone might not be enough to be distinguishable for everyone, Microsoft (n.d. a) suggests using a colour accessibility checking tool.

The last category includes some basic principles derived from the guideline of the University of Washington to be followed when designing a questionnaire (University of Washington, n.d.). Clear instructions on how to achieve this are provided by a checklist at the end of the design. To conclude, all these principles derived from various guidelines aim at helping to design an accessible PDF. That way, they form the basis for major design choices.

Table 2*Design Principles Derived and Implemented in the Different Prototypes*

Principles Considered in the Design	Prototype 1	Prototype 2
1. Media		
<i>Implemented by:</i>		
a) Visual response alternatives	YES	NO
b) Alternate text to images	YES	YES
c) Minimize elements that can cause visual distractions	NO	YES
2. Text		
<i>Implemented by:</i>		
a) Font size of 16pt or bigger	YES	NO
b) Bigger line spacing	YES	NO
c) Descriptive and clear pages and titles for the instructions	YES	YES
d) Recommended font type	YES	YES
e) Break up large amount of text into smaller paragraphs	YES	YES
3. Navigation and Links		
<i>Implemented by:</i>		
a) Provide additional explanations for clarity	YES	YES

b) Gestalt principles (Proximity & Similarity)	YES	YES
c) Radio buttons instead of typical table	YES	NO
d) Conveying information through multiple medium	YES	YES
e) Use built-in headings	YES	NO

4. Colours

Implemented by:

a) Accessible colours	NO	YES
b) Contrasting colours	YES	YES

5. Assistive technologies

YES YES

Implemented by:

a) Identify document language	YES	YES
b) Compatible with assistive technology	YES	YES
c) Use tables wisely	YES	NO
d) Export Word into barrier-free PDF	YES	YES

Note. YES means that the corresponding criterion was met by the prototype version, NO indicates that the criterion stated was not considered in the design.

Based on the principles (Table 2), design choices were made in two different versions of the initial redesign of the BUS, resulting in two prototypes. For the first prototype, two out of three principles for the category media were implemented. In Figure 1, principle 1a) (Table 2) was implemented by including smileys as visual response alternatives. In addition, each pictorial was described with alternate text (Principle 1b), Table 2). For the second prototype, only principle 1c) was implemented visibly in the design. To reduce visual distractions, no visual response alternatives or other forms of pictorials were implemented (Figure 2).

Figure 1

Design of an example question of Prototype 1

1. The chatbot function was easily detectable. 🔍

1 **☹️** 2 **😐** 3 **😐** 4 **😊** 5 **😄**

STRONGLY DISAGREE DISAGREE NEITHER DISAGREE NOR AGREE AGREE STRONGLY AGREE

Figure 2

Design of an example question of Prototype 2

1. The chatbot function was easily detectable.

1 2 3 4 5

Strongly Disagree Neutral Strongly Agree

The second category of principles refers to Text. For prototype 1, principle 2a), a font size of 16, principle 2b) using line spacing of 1.5 and 2d) using Arial as the font type can be seen in Figure 1. The other two principles (2c), 2e), and Table 2) were also implemented by


providing clear instructions and titles and breaking up the instructions into smaller paragraphs (Figure 3).

For the second prototype, three of the Text principles were followed. Thereby, Arial was also used as a recommended font type (Figure 2; Principle 2d), Table 2). 2c) Descriptive and clear pages and titles for the instructions were combined with 2e) breaking up those instructions into smaller parts (Figure 4), even though this was not as clearly separated as in prototype 1.

Figure 3

Design of an instruction part of Prototype 1

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:

Figure 4

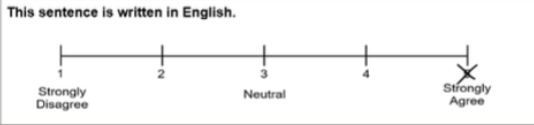
Design of an instruction part of Prototype 2

Section one instructions:

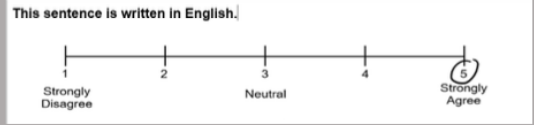
In this section you are given three sentences. Choose only one of the possible answers on how much you disagree or agree with the sentence. Put **X** or **O** over the number of your choice.

This is an exemplary sentence with an answer:

This sentence is written in English.



This sentence is written in English.



All five principles under 3. Navigation and Links were considered when designing prototype 1. Additional explanations were given by explaining all steps needed to fill in the questionnaire (Figure 3; Principle 3a), Table 2). In addition, c) buttons were incorporated as a means to answer, d) information was conveyed through multiple mediums by adding pictorials as response alternatives, but also by adding additional pictorials and e) different levels of built-in Word headings were utilized for the different parts (Figure 1). Relating to principle 3b), Gestalt principles of proximity and similarity were followed. Proximity can be seen by keeping two questions per page and by separating them with lines in between (Figure 5). The similarity was implemented by keeping the design of each question the same, demonstrating that these questions together belong to the questionnaire (Figure 5).

Figure 5

Example page extracted from prototype 1

2. It was easy to find the chatbot. 🔍

1 2 3 4 5

☹️ ☹️ ☹️ 😊 😊

STRONGLY DISAGREE DISAGREE NEITHER DISAGREE NOR AGREE AGREE STRONGLY AGREE

3. Communicating with the chatbot was clear. ✓

1 2 3 4 5

☹️ ☹️ ☹️ 😊 😊

STRONGLY DISAGREE DISAGREE NEITHER DISAGREE NOR AGREE AGREE STRONGLY AGREE

4

In the design choices for prototype 2, principle a) additional explanations were also implemented by providing instructions and further cues (Figure 4, Figure 6). Gestalt

principles of proximity and similarity were included following the same logic as in prototype 1, but additionally, instructions were coloured differently to stress similarity more (Figure 7). As the last point from this category, prototype 2 corresponds to principle 3d) conveying information through multiple mediums by using the lines corresponding to each value of the Likert scale, in combination with words (Figure 7).

Figure 6

Additional instructions on the bottom of the page from prototype 2

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.

Figure 7

Example Layout of Prototype 2

<p style="text-align: center; margin: 0;">Section two instructions:</p> <p style="font-size: small; margin: 0;">You will be given 11 statements to read. For each statement choose <u>only one</u> of the possible answers on how much you disagree or agree with the sentence. Put X or O over the number of your choice.</p> <p>1. The chatbot function was easily detectable.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>2. It was easy to find the chatbot.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>3. Communicating with the chatbot was clear.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>4. The chatbot was able to keep track of context.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>5. The chatbot's responses were easy to understand.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p style="text-align: center; margin-top: 10px;">Continue to next page</p>	<p style="text-align: center; margin: 0;">Continuation of previous page</p> <p>6. I find that the chatbot understands what I want and helps me achieve my goal.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>7. The chatbot gives me the appropriate amount of information.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>8. The chatbot gives me the information I need.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>9. I feel like the chatbot's response was accurate.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p>10. I believe the chatbot informs me of any possible privacy issues.</p> <p style="text-align: center;"> 1 2 3 4 5 Strongly Disagree Neutral Strongly Agree </p> <p style="text-align: center; margin-top: 10px;">Continue to next page</p>
--	--

For both prototypes contrasting colours were used, as can be seen in all examples (Principle 4b), Table 2). Accessible colours, in line with W3C standards, were only used in

Prototype 2, considering that only black and grey were used (Principle 4a), Table 2). For prototype 1, the colours were not checked for accessibility.

The last category of principles, 5. Assistive technologies do relate more to the process of designing rather than visible design choices. In both prototypes all the principles were followed during the design, using the accessibility instructions provided by the University of Washington (University of Washington, n.d.). Only principle 5c) was not followed specifically in prototype 2. Otherwise, both prototypes did not show any accessibility issues.

Taking the different principles into consideration, it resulted in two prototypes of the BUS-A. To explore which design is more accessible or if the two designs should be merged a focus group was performed in the next step.

Focus Group About the BUS-A Prototypes

Participants

The focus group was conducted in collaboration with the Open Mind School in California which supported the recruitment of participants with disabilities. The questionnaire design and data collection were also done in partnership with a further researcher, Maria Hristova. All data was shared between both researchers. Participants in the focus group tested the questionnaire by evaluating a chatbot with a pre-specified task they had to conduct first. Before continuing with the second phase of the study, a final version, the BUS-A was designed. Ethical approval was given for both phases of the study by the Ethics Committee of the Faculty of Behavioural Sciences (Request number 220273).

For the focus group, the Open Mind School recruited two participants via convenience sampling to evaluate the prototype versions of the redesigned BUS. Both participants were male, aged 17 and 21 ($M = 19$). Both participants in the focus group had autism and a communication assistant was present to help with the understanding of the interview. Inclusion criteria were to have a disability and to be aged 15 or up. Informed consent

(Appendix A) was signed by the students themselves before the focus group, as they were already older than 16.

Materials

For the focus group, a protocol was agreed on first, containing all relevant information on the aim of the session and important probes to be asked (Appendix B). Another important material prepared for the focus group was a separate informed consent sheet that needed to be signed. In addition, a YouTube video explaining chatbots in general terms was also used during the focus group (<https://www.youtube.com/watch?v=pX6zqaEHAdw>).

A task (Appendix C) was given to be conducted using the chatbot of Zoom. The task concerned finding information on how to change the virtual background and the link to the Zoom website (<https://zoom.us/>) was already provided in the task. To conduct the task, a laptop was provided for each participant. Two printed prototypes of the BUS-A (Appendix D & Appendix E) were supplied to be filled in by the participants. Both versions contained the 11 items of the BUS, which can be answered on a five-point Likert scale, ranging from “strongly disagree” to “strongly agree”. Both versions contained a demographic questionnaire, asking about age, gender and type of disabilities.

For understandability of the interview, a communication assistant was present to help the students. The focus group resulted in a recording and a transcription of the session (Appendix F), serving as a basis for the redesign. Materials utilized for the redesign of the BUS-A were Word for the paper version and Qualtrics (Qualtrics, n.d.) for an accessible online version.

Procedure

The procedure (see Appendix B) of the focus group was agreed upon between the team of Open Mind and the one of the University of Twente by teleconference before the actual start of the focus group. One representative of the Open Mind School received all the information on the study’s needs and recruited suitable participants that way. The

representative conducting the focus group collected informed consent to participate in the study before actually starting.

The participants were first introduced to the goal of the study and the topic of chatbots by showing a short YouTube clip about them. After finishing the task on the Zoom chatbot they were provided with the two prototype versions of the BUS-A on paper. The students filled in both questionnaires quietly on their own. Next, the open discussion about the questionnaires started and students answered the probes with help of the communication assistant who rephrased the interview questions for them. The focus group took about 30 minutes in sum.

After the actual focus group, the researcher of the Open Mind School provided the recording and transcription of the interview. These served as a basis for redesign, which was conducted in Word and Qualtrics (Qualtrics, n.d.). In addition, the Word version was converted into an accessible PDF that could also be printed.

Data Analysis

The audio recording and transcription of the focus group were used to extrapolate comments and issues regarding the usage of the two prototypes of BUS-A. This was done by listing all identified positive and negative points mentioned during the focus group in a table, sorting them per element of the prototype, based on pre-specified principles (Table 3). Finally, issues associated with the compression of the text and the design of the scale were listed and considered to be implemented in the final version of the BUS-A.

Results

The participants of the focus group highlighted four main points to improve the accessibility and usability of the BUS-A (Table 3). First, both participants indicated that they did not like the font size. After probing for further details, they both mentioned that they would like to have it bigger (Table 3, point i)). One of the participants answered liking the smileys as an additional answering option: “Did you like or dislike it [the smileys]?” – “Liked

it [the smileys]”. Regarding point iii), both participants approved of the use of colours in prototype 1, which was not implemented in the second prototype. This was one of the reasons both participants preferred the first prototype overall: “Did you like the first one [prototype 2] or second one [prototype 1] more?” – “Second one [prototype 1]. Lastly, one of the participants indicated that they did not like the question about familiarity with the chatbots (Table 3, point iv)). Based on these main points based on feedback from the focus group, a redesign was conducted.

Table 3

Elements in the Prototypes Commented on by the Participants of the Focus Group based on Principles pre-specified in the design process (Table 2)

Elements commented on in the focus Group	Prototype 1		Prototype 2	
	Participant 1	Participant 2	Participant 1	Participant 2
i) Font size (as suggested by <i>principle 2a</i>)	Criticized	Criticized	Criticized	Criticized
ii) Smileys as visual response alternatives (as suggested by <i>principle 1a</i>)	Approved	-	Criticized (as not implemented)	-
iii) Use of colours	Approved	Approved	Criticized (as not implemented)	Criticized (as not implemented)
iv) Questions about the familiarity with chatbots	-	Criticized	-	Criticized

Total Criticized/ Approved	3 points of approval	/ 3 points of criticism	0 points of approval	/5 points of criticism
-------------------------------	-------------------------	----------------------------	-------------------------	---------------------------

Note. Elements that were commented on by each of the two participants separately, denoted by *Approved* when participants explicitly mentioned that they liked the element and *Criticized* for when they actively reported issues associated with the element.

Discussion Phase 1

Based on insights from the focus group, the final version of the scale could be designed (BUS-A; Appendix G). The final version was designed by working on prototype 1 and making changes in this document, as the participants of the focus group pointed out that they preferred this design over the other and as it already performed better on points of critique of the first prototype (Table 3).

To make it more accessible, the overall colour scheme was changed based on accessibility standards (Color Safe, n.d.). These standards ensure that the colour contrast is accessible according to WCAG Guidelines. For each different heading level, a different variation of blue was chosen. Additionally, the smileys were coloured in this redesign, as the students pointed out that they liked the colourful design. Again, these colours were based on accessibility standards.

As requested by the participants of the focus group, the font size was made bigger. Namely, 18pt. Next to changes in the colour scheme and overall format, some changes in instructions were made and some details were added. Instructions for respondents on how to fill in the questionnaire were placed in more fitting places to increase understandability. That way, aiming at a clearer structure. For the demographics, the questions on the category of disability respondents could select were rephrased and the gender question was made more sensitive.

Lastly, two new elements were added to the questionnaire. For the purpose of data analysis, a field to fill in the number of participants was included on top of each page. This is especially crucial for the paper version of the questionnaire, ensuring that each page is connected to a participant, in case some pages get lost. Also, a title page was added to the questionnaire, stating the name of the scale and the researchers and institutions involved in creating it. Additionally, an explanatory note for the filling in of the participant number was placed on the title page as a disclaimer.

After converting the BUS-A from Word into an accessible PDF (Appendix G), the scale was also set up in Qualtrics (Qualtrics, n.d.), using the same design choices as in Word. That way it ensures more reliable and more convenient data collection. Now, students could also decide for themselves whether they want to use the paper version or the digital version of the survey. The digital version also included consent and the task description. Thus, participants got the chance to conduct the study themselves at home.

Testing BUS -A and Comparative Analysis with BUS-11

Methods

In the second phase of the study, a between-subject design was used by asking participants with and without disabilities to interact with the Zoom chatbot and then assess their satisfaction with BUS-A or BUS-11. An experimental group with participants with and without disabilities assessed their satisfaction using the BUS-A, the control group used the BUS-11 and it was composed only of participants without disabilities. Participants without disabilities were randomly assigned to the experimental or control group.

Participants

First, missing data were excluded. Altogether, 116 participants aged between 15 and 60 ended up taking part in the study ($M = 24.87$, $SD = 8.28$). 70 participants were female, and 46 were male. One participant was transgender, one indicated being genderqueer, and two participants were non-binary. Participants were recruited using convenience sampling for the

control group, and purposive sampling for the experimental group. For the control group, the study was uploaded to the Test Subject Pool System (SONA) of the University of Twente (Utwente Sona Systems, n.d.), through which students can gain credits by participating in studies. In addition, snowball sampling was used asking acquaintances of the researchers to participate. For the experimental group, participants were recruited via several means of convenience sampling – Facebook, Sona Systems and different Forums were joined to reach as many participants as possible. Lastly, the study was distributed on Facebook by joining groups specifically aimed at exchanging surveys.

Inclusion criteria were proficient English skills. The control group was composed of 90 participants, ages ranging from 15 to 54 (Mean age = 24.19, SD = 7.63). 52 participants were female, and 41 were male. 1 participant was transgender, 1 Genderqueer and 2 participants indicated being non-binary. Informed consent was provided online before asking for demographics.

For the experimental group, participants were also recruited via convenience sampling, but efforts were also made to recruit the target group via different forums for disabled people. In the end, the study was conducted by 26 participants aged between 21 and 60 (18 female, Mean age = 27.31, SD = 10.07). About 19% of the participants reported having one or more disabilities. Out of those, the types of disabilities declared by the respondents were approximate: 14% developmental disability, 28% mental health or emotional disability, 28% unseen disability, 14% physical disability and 14% sensory disability. Before collecting data, ethical approval was given by the Ethics Committee of the Faculty of Behavioural Sciences.

Materials

The data collection in phase two consisted of two parts: i) data collection with the experimental group, and ii) data collection with the control group. Each part included shared, but also separate materials. Overall, both groups conducted the study on a laptop or a mobile device using the software Qualtrics (Qualtrics, n.d.). Before the actual questionnaire, an

introduction to the study and a demographic questionnaire was provided. Afterwards, both groups received the same task about changing backgrounds, corresponding to the chatbot of Zoom (<https://zoom.us/>). What differed between the two groups was the type of questionnaire to be filled in: i) the experimental group received the BUS-A, ii) the control group received the BUS-11. Although presented in a different design, both questionnaires entail 11 items, to be answered on a five-point Likert scale, ranging from “1 – Strongly Disagree” to “5 – Strongly Agree”. Excel and RStudio were then used for data analysis.

Procedure

Prior to the start of the study, ethical approval was requested by the Ethics Committee. The research project request was approved on March 18th, 2022. After evaluating the results of the first phase of the study, a final version of the BUS-A was constructed together with the second researcher. The main part of the data collection consisted of two parts: i) the control group, ii) the experimental group. People with disabilities were automatically allocated to the BUS-A, and people without disabilities were randomly allocated to either BUS-A or BUS-11. Overall, the procedure for both of the groups was the same. After being informed about the study and giving informed consent, the participants were introduced to demographical questions. They were also informed that they could withdraw from the study at any time without further explanations. Next, they received their chatbot and the corresponding task. After conducting the task, they were asked to come back to Qualtrics (Qualtrics, n.d.) and fill in the questionnaire. That way, completing the study.

Data Analysis

Data was exported from Qualtrics (Qualtrics, n.d.) as a CSV file. Before being imported into R Studio (Rstudio, n.d.), the data set was prepared in Excel (Microsoft, n.d. b). Responses with non-sufficient survey progress were deleted, and unnecessary columns of data were also excluded from the data set. From that final data set that included all the data (D1), two separate data sets were also created: i) one containing only the data of participants in the

control group who performed their assessment of the chatbot filling in the original version of the scale BUS (DBUS), and ii) one containing the data of participants with and without disabilities of the experimental group who assessed the chatbot using the BUS-A (DBUSA).

Analysis was conducted in RStudio (Rstudio, n.d.) (Appendix H). All item scores were transformed into percentages to account for the Likert Scale used. Descriptive statistics and frequency analysis were performed to explore individual characteristics of the participants in the control and experimental groups such as age, gender, and type of disabilities. Medians and standard deviations were calculated. To test whether the data is distributed normally, a Shapiro Test was run, and QQ-Plots were constructed using the data from D1 (Yap & Sim, 2011).

Considering the research question on whether the two versions of the scales show any significant difference in terms of overall satisfaction assessment of the chatbot of the participants in the two conditions (control and experimental), a Welch two-sample t-test was performed. To ascertain that the two versions of the scale did not result in a significantly different level of satisfaction, analysis was performed using the overall scores of satisfaction obtained using BUS-11 and BUS-A. Moreover, because the experimental group was composed of people with and without disabilities while the control group was composed of participants without disabilities, to check if people with disabilities rated their satisfaction in a significantly different way using the BUS-A compared to people without a disability using both BUS-A and BUS-11, a linear regression model was implemented. Also, a Welch two-sample t-test was performed to test for differences between people with and without disabilities using the BUS-A. Furthermore, to investigate whether the factorial structure of the two versions of the scales differs from the expected 5 factors identified by Borsci et al (2021b), a confirmatory factor analysis (CFA) was performed using the R package lavaan and semPlot for visualisations (Rstudio, n.d.).

The number of questionnaires collected could be considered below the threshold for a reliable CFA (Comrey & Lee, 1992), especially when DBUS and DBUSA are investigated separately. In this context, we used CFA as a way to qualitatively observe differences between the original and the accessible version of the scale. To assess the overall fit in a CFA, experts are usually looking at multiple criteria (Pavlov et al., 2020; Sivo et al., 2006). First, attention had to be paid to Chi-square, which is supposed to be significant to indicate a good fit of the five-factor model. Next, the comparative fit index (CFI) was analysed, which should be around .95 or bigger to indicate a good fit (Sivo et al., 2006). Further two indexes important to be investigated in a CFA are the Root Mean Square Error of Approximation (RMSEA), actually testing the model misfit (Pavlov et al., 2020), which needs a value of $<.06$ to indicate an absolute, good fit (Sivo et al., 2006) and the Standardized Root Mean Square Residual (SRMSR), also testing the model misfit, indicating a good fit when $<.7$ (Pavlov et al., 2020). Finally, the reliability of the two versions of the scale, in addition to the overall dataset, was monitored by testing item-total correlation and Cronbach's alpha, aiming for a level of reliability of the scales. A score over 0.8 would be considered to be in tune with the validation of the BUS scale (Borsci et al., 2021b).

Results

Descriptive Statistics

Before starting with the data analysis, all item scores were transformed into percentages. That way, accounting for the Likert-Scale. After examining frequencies on participants, descriptives for dataset 1 were run. Means and Standard Deviations of satisfaction measured by both versions of the BUS can be seen in Table 3. Item 10 deviates most from the average ($M = 48.45$, $SD = 20.96$), all other items seem to be distributed around the average mean ($M = 80.08$, $SD = 12.01$).

Table 3

Median Satisfaction Level per Each Item and on Average in Percentages

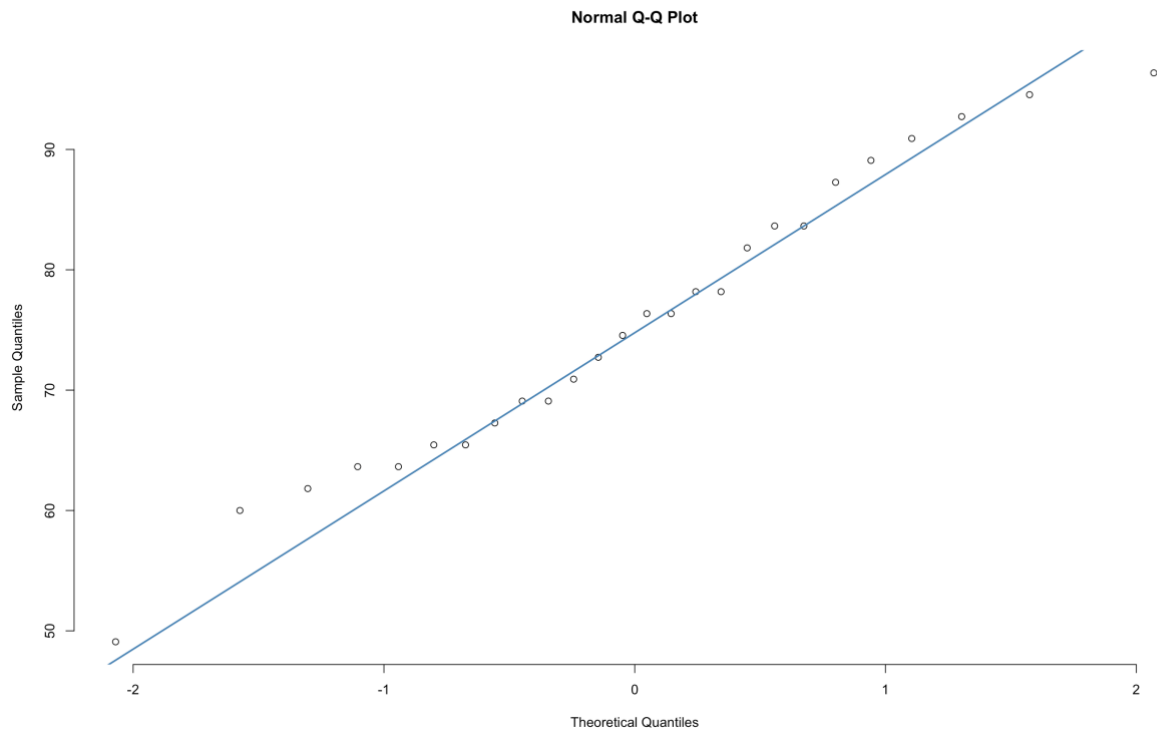
BUS Item Number	Mean	Standard Deviation
Item 1	83.97	19.42
Item 2	85.34	19.62
Item 3	84.14	17.40
Item 4	80.00	19.02
Item 5	86.72	15.37
Item 6	81.55	19.23
Item 7	81.21	19.96
Item 8	75.69	21.07
Item 9	84.14	17.20
Item 10	48.45	20.96
Item 11	89.66	14.80
BUS Average	80.08	12.01

To test for normality, two methods were followed. First, a Shapiro-Wilk test was conducted on each item of satisfaction in D1. All items were not normally distributed, considering their significant p-value of $<.05$. For the average score of the BUS, the Shapiro-Wilk Test resulted in $W = 0.952$, $p < .001$. Thereby, again indicating that the data is not normally distributed.

In the second step, a Quantile-Quantile (Q-Q) plot was run on the overall score of each item to visualise the distribution of the data. Figure 8 indicates that the overall score in percentages is normally distributed. Still, the Shapiro test resulted in a non-normal distribution. Meaning that non-parametric tests should be used in the following analyses.

Figure 8

Q-Q Plot for the Average Overall Satisfaction Score in Percentages



Hypothesis Testing

To investigate whether the two versions of the BUS differ significantly from each other in terms of overall satisfaction scores reported (RQ1), a Welch two-sample t-test was performed between the mean scores of the BUS-11 and the BUS-A, but also between each factor of the scales. Results show a significant difference between the mean score of the BUS-11 scale ($M = 81.414$) and the one of the BUS-A ($M = 75.455$), $t(39.682) = -2.235$, $p = .031$. This indicates significantly different levels of satisfaction when using the different scales. Furthermore, the analysis of the average scores of each factor did not highlight any significant differences in three of the five factors (Table 4). However, factor 2 and factor 3 indicate a slightly significant difference between both groups, as shown by the significant p-values. Therefore, creating the need to further investigate whether a difference is striking.

Table 4

T-tests Between Factors of DBUS and DBUSA

Factor	T	Df	p	M Control (DBUS)	M Experimental (DBUSA)
--------	-----	------	-----	-----------------------	-----------------------------

Factor 1	-0.252	47.213	0.802	84.444	85.385
Factor 2	2.190	114	0.031**	85.259	77.949
Factor 3	2.709	37.781	0.010**	82.833	73.077
Factor 4	1.730	52.397	0.090	50.000	43.077
Factor 5	-0.132	39.58	0.896	89.556	90.000

** $p < 0.5$

The regression analysis also performed to answer the first research question regarding differences in satisfaction scores, between satisfaction reported by people with disabilities using BUS-A and the people without disabilities using the BUS-A or BUS-11 suggested that there is no significant difference due to the combination of the type of participants and scale used for the assessment ($F(2, 113) = 2.643, p = 0.076$), with an $R^2 = 0.028$.

The second research question specifically aimed at assessing differences between people with and without disabilities conducting the BUS-A. It was tested by performing a Welch two-sample t-test between the mean scores of people with disabilities who used the BUS-A ($M = 73.818$) and people without disabilities who used the BUS-A ($M = 75.844$). There was no significant difference between satisfaction scores based on whether one had a diagnosed disability ($t(5.070) = -0.272, p = .796$).

Psychometric Assessment of the Scales

To answer the third research question of whether the five-factor structure of the BUS-11 can be confirmed for the BUS-A, a CFA was computed using the pre-specified structure (Borsci et al., 2021b). Considering the problematic distribution of the data, ML robust was used for all factor analyses. Factor loadings and further indexes were compared between the BUS-11 (CFI = .947, RMSEA = .085, SRMSR = .053) and the BUS-A (CFI = .594, RMSEA = .229, SRMSR = .131). The Chi-square of BUS-11 resulted in a good fit, and Chi-square of the BUS-A also in a good fit (Table 5). Thereby, differences in fit indexes indicate some disparities between the two versions of the scale. Generally, the BUS-A shows a bad model fit

for the suggested five-factor structure (Table 5). For the BUS-11, only RMSEA does show a misfit.

Table 5

Goodness of fit Model of the Five-Factor Structure for Satisfaction

Scale	χ^2	df	<i>N</i>	<i>p</i>	CFI	RMSEA	SRMSR
BUS-11	59.295	36	90	0.009	0.947	0.085	0.053
BUS-A	105.397	36	26	<0.001	0.594	0.229	0.131

Moreover, as can be seen in Figure 9 and Figure 10, there are some differences between the two scales in factor loadings of factors two and three. More specifically, item 5 and item 8 seem problematic. Although these factors show rather low loadings for the experimental group, the same pattern of factor loadings is reflected in both of the scales. Furthermore, a difference gets evident in the correlations between the separate factors themselves. As can be seen in Figure 9, a small but negative correlation appears between factor 4 and factor 5 in the experimental group. This negative correlation is neither evident for the control group (Figure 10), nor for the complete dataset (Figure 11). This implicates that there might be a relationship between perceived privacy and security and time the chatbot takes to answer. Despite the BUS-A indicating a bad model fit, except for Chi-square, factor loadings thus seem in line with the ones of the BUS-11. Therefore, scales will be treated as equivalent and further analyses will be performed on the whole dataset (D1).

Figure 9

Factor Loadings Experimental Group

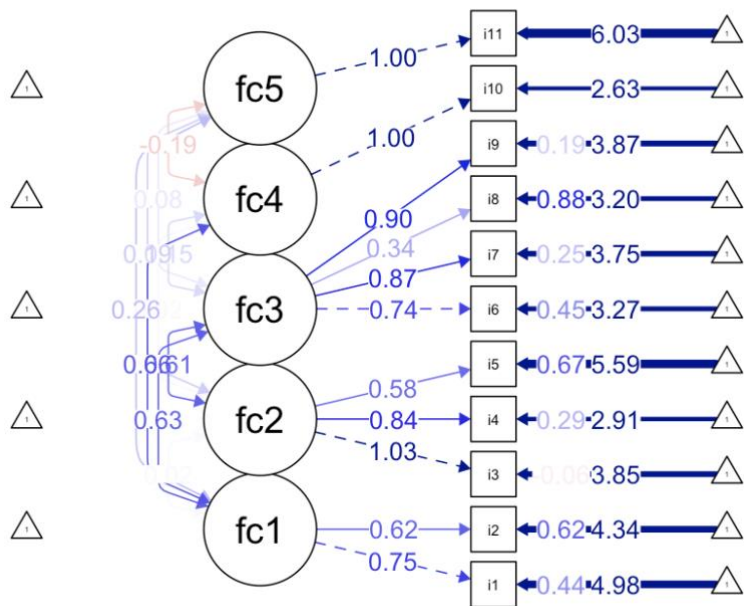
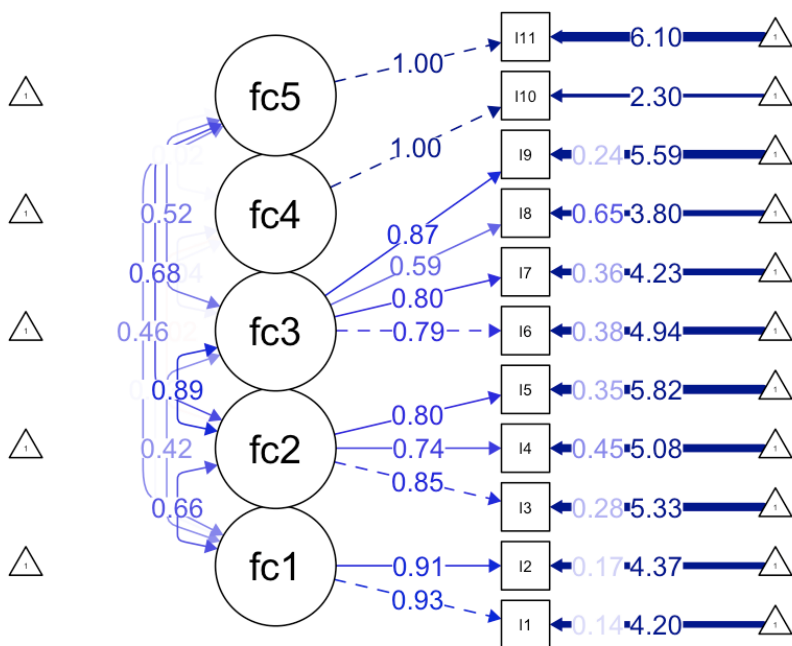


Figure 10

Factor Loadings Control Group



Using the entire dataset (D1) for the final CFA, the results showed an overall good fit of the five-factor model, as validated by the robust indexes (CFI = .941, RMSEA = .080, SRMSR = .048). In addition, Chi-square also resulted in a good fit (Table 6). Thus, CFA validated the initial five-factor model for both scales combined.

Table 6

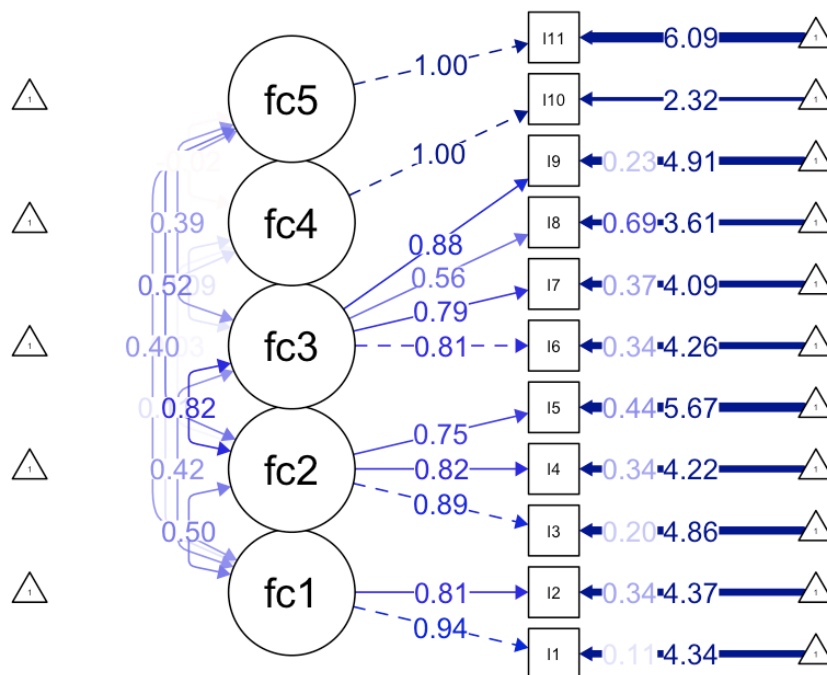
Goodness of fit Model of the Five-Factor Structure for Satisfaction for the Overall Data

Scale	χ^2	df	N	p	CFI	RMSEA	SRMSR
Combined data	62.576	36	116	0.004	0.941	0.080	0.048

Moreover, factor loadings (Figure 11) show the same pattern as was expected. Factor loadings are high for almost every item, except for item 8, which in contrast to the other loadings only displays a loading of 0.56. In combination with the fit indexes, the model can be concluded to have a good fit overall.

Figure 11

Factor Loadings of Complete Dataset



Reliability Analysis of the Different Versions of the Scale

As a means to assess the quality of the items of the BUS-A and to compare it to the quality of the original scale (RQ3), the item-total correlation was computed for each of the items (Table 7). Therefore, r.drop is reported (Table 7). R.drop indicates the item-total

correlation without the item itself and is considered as correlating well with the scale if being high (Rpubs, 2015).

For the overall data (D1) of the BUS, item 10 does not seem to correlate with the overall scale. Otherwise, all items reflect to be correlated with the scale, demonstrating a good quality of the items. For the BUS-11, the same pattern is reflected with only item 10 being not correlated at all with the scale. For the BUS-A, three items do not correlate as strongly with the scale: Item 2, Item 10 and Item 11. Altogether, these items are also not as strongly correlated as other items in all of the datasets.

Table 7

Item-total correlations per each item for BUS overall, the BUS-11 only and the BUS-A

BUS Item Number	r.drop BUS	r.drop BUS-11	r.drop BUS-A
Item 1	0.54	0.59	0.46
Item 2	0.45	0.52	0.26
Item 3	0.74	0.76	0.70
Item 4	0.68	0.67	0.71
Item 5	0.71	0.75	0.50
Item 6	0.69	0.70	0.60
Item 7	0.65	0.65	0.64
Item 8	0.47	0.42	0.52
Item 9	0.75	0.76	0.69
Item 10	0.09	0.04	0.18
Item 11	0.48	0.58	0.20

Additionally, Cronbach's alpha was analysed to assess the reliability of the whole scale, but also of three of the factors separately. The reliability of the whole scale can be

considered as high ($\alpha = .859$). Separate factors show similar reliability indexes (Table 8).

Thus, the scale overall can be considered reliable, just as every one of the factors.

Next to the reliability of the overall scale, the separate scales were also tested for Cronbach's alpha. Both scales, the BUS-11 ($\alpha = .863$) and the BUS-A ($\alpha = .830$) show high reliability, just as each of the factors (Table 8). Not in tune with the reliability coefficients found by Borsci et al. (2021b) is Factor 1 of the BUS-A ($\alpha = .632$), which is lower than for the other scales. Factor 3 ($\alpha = .785$) also displays lower reliability than aimed at but is close to the cut-off score of 0.8. Still, all data can be considered reliable (RQ3).

Table 8

Cronbach's Alpha per Factor and Overall for BUS overall, the BUS-11 only and the BUS-A

Factor	Cronbach's alpha		
	BUS	BUS-11	BUS-A
Factor 1	0.906	0.915	0.632
Factor 2	0.837	0.841	0.875
Factor 3	0.855	0.845	0.785
Overall	0.859	0.863	0.830

Discussion Phase 2

The second phase of the study aimed to test the BUS-A and compare its results to the outcomes of the BUS-11. Generally, results indicated a significant difference between the two scales, the BUS-A and the BUS-11, but not between the three groups conducting the questionnaires (RQ1). This implicates that the reason for the difference the t-test showed might lie in the assessment instruments, not in having diagnosed a disability or not. Regarding the second research question and in tune with accessibility research, no difference was found between people with and without disabilities conducting the BUS-A. Thereby, demonstrating that having a disability does not affect satisfaction levels.

All groups were able to fill in the questionnaire. This again implies that the two scales themselves might lead to the different levels of satisfaction reported. On the one side, this implies that the BUS-A is accessible. Thus, showing that principles followed have actually increased the “range of users able to operate” (Moreno et al., 2012), which is described as the key to accessibility. On the other side, it points out that the design might have had an effect on the different satisfaction scores but not the accessibility aspect itself. Although Schrepp et al. (2016) suggested that adaptations to the design need to be made to ensure accessible research, there was no specific guide. Following different principles found in an initial literature review did not lead to a design with the same outcomes as the original scale, although being accessible. Thereby, implicating that principles regarding inclusive design might not be specific enough, nor validated enough.

Furthermore, a comparison between the factorial structure of BUS-A and BUS-11 showed some slight differences in some of the items but a significant difference in terms of model fit. Certainly, the very low fit of the BUS-A could be attributed to the sample size. Nevertheless, the low fit also indicates that the BUS-A might actually be a significantly different assessment instrument compared to the original BUS-11. This was also reflected in the reliability analysis, where although being reliable overall as a scale, the factors of the BUS-A did not show the same reliability. Additionally, a slight negative correlation was found in the BUS-A between perceived privacy and security (Factor 4) and time response (Factor 5). This was not found for the BUS-11 and should thus be further investigated. Although it only shows a slight correlation, that does not point towards causation. Given the general factorial differences in the model fit between the BUS-A and BUS-11, this correlation needs to be tested in future research. This could help investigate whether the BUS-A is indeed a different testing instrument.

These differences regarding the factorial analysis and the reliability of the factors could be due to the new design elements added in the first part of the study. Although design

principles and standards were followed, there is no actual, clear guide on questionnaire design, leaving certain decisions to the researcher. In line with this, Cernat and Liu (2018) found that smileys, which were incorporated into the design of the BUS-A, might lead to worse quality of the data collected. Thereby, influencing the data collected and the corresponding results. Despite the difference in the two scales, the CFA carried out on the entire dataset suggests that the five-factor structure of the BUS-11 (Borsci et al. 2021b) is overall respected, also including the BUS-A (RQ3). Future studies should investigate whether the bad model fit of the BUS-A can be replicated or whether it is due to the low response rate.

This part of the study aimed at focusing not only on usability but also on accessibility (Walmsley, 2001), which is an important step in developing inclusive technologies for disabled people, thereby opening up new possibilities for them (Tsatsou, 2020). Results implicate that both versions of the scale are valid in testing usability of chatbots, also serving as a general validation of the BUS into usability research.

Overall Discussion

The current study aimed at creating a tool that is accessible for everyone based on the BUS-11, to then test and validate it and to also validate the original scale in the last step. Therefore, two research questions aimed at testing the BUS-A, which was created in the first step. It was analysed whether the results of the two versions of the scale significantly differ from each other and whether having a disability had an effect on the results. In the last step, the factorial structure and reliability of the two scales were analysed.

The first part of the study resulted in an accessible version of the BUS-11, which was validated regarding its accessibility in a second step. In line with research, people with disabilities contributed to the process by testing the prototypes and commenting on them (Walmsley, 2001). Important principles were pointed out that needed special consideration in accessible design: the importance of bigger font size, the use of colours and smileys as response alternatives. As two prototypes were created, this feedback gave important insights

on which prototype to prefer. Generally, a focus group led to better-informed design choices (Barrett & Kirk, 2000).

The principles identified as especially favourable by participants of the focus group are in line with previous research. A bigger font size was specified by most of the literature, but specific size varied between the advices. The use of colour contrasts is supposed to increase readability (Vanderheiden et al., 2021). Although the second prototype also used contrasting colours by definition, black and white seemed too monotone. Actually, colours can be used as a tool to create structure and give a guideline by evoking different associations and feelings (Sik Lanyi, 2017). Another important implication is the use of smileys as a visual response alternative. This design recommendation was not included in most of the literature but favoured by both participants of the focus group. As suggested by Hartley and MacLean (2006) this helps in understanding the answer options of the Likert Scale. The principle that if a product would be inclusive for people with disabilities, it should also be easily operable for people without disabilities was followed throughout the study design (Sik Lanyi, 2017).

The research question “*Are the two scales, BUS-A and BUS-11 significantly different from each other in terms of participants' satisfaction rate with a chatbot?*” can be answered with yes, which is not in line with expectations. First, analyses showed a significant difference in satisfaction levels between the BUS-11 and the BUS-A. This was not in tune with literature, which pointed out that by implementing principles of accessibility, scales can be made more inclusive, but should not have an influence on satisfaction levels (Schrepp et al., 2016). A linear regression then resulted in no significant differences between the experimental groups. Thereby, pointing out a discrepancy between the two versions of the questionnaire, but not between the people conducting it. This discrepancy could be due to the difference in the two assessment measurements, not due to an effect of the people and their characteristics conducting them.

So, a possible explanation for the unexpected outcomes regarding the first research question could be that the scales themselves differ in their structure, although people were all able to conduct the questionnaires the same way, no matter whether having a disability or not. As has been pointed out in research, adaptations for inclusive design are often rare (Harper & Chen, 2011). Therefore, the design principles followed could have been not validated enough, leading to the unexpected outcome. When design principles are only implemented in 10% of the cases (Harper & Chen, 2011), they cannot be tested enough to reveal issues with them. Thus, although design recommendations were systematically followed and even tested, those might have changed something in the underlying structure of the questionnaire, which will get evident when taking a closer look at the factorial analysis.

The second research question was: “*Do people with and without disabilities report significantly different satisfaction scores using the BUS-A?*”. In line with research and expectations, no significant differences between satisfaction scores of people with and without disabilities conducting the BUS-A could be found. That means that both groups of people were able to fill in the questionnaire and that both groups of people ended up with similar satisfaction scores. This implies that the scale itself can be considered accessible and inclusive for people with and without disabilities.

In tune with research, it was found that specific adjustments in setting, timing, presentation and response did finally lead to an accessible version of the scale (Goegan, 2018). This was replicated by incorporating different design principles in the current study. When relating these outcomes to the outcomes of RQ1, an explanation for the discrepancy between the two test results could be either due to sample size or another effect that would need to be investigated. Again, a possible explanation would be that the scale itself can be considered accessible, but not the same as the original scale in terms of the underlying structure. Considering the very limited sample size of disabled participants, it would be advisable to re-test it with a bigger population.

The third research question “*Are the psychometric properties of the BUS-A (factorial structure and reliability) in line with the one of the original scale?*” has partly already been answered by relating to differences between the BUS-A and BUS-11. Other than expected, the factorial structure could thus not be confirmed. Outcomes of the analysis show a bad model fit. Unlike expected, outcomes implicate that two different assessment instruments need to be distinguished. The BUS-A does not fit with the originally validated model of the BUS-11. Also, a negative correlation between two of the factors, factor 4 and factor 5, appears for the BUS-A which was not found for the BUS-11. Thereby, future studies would need to analyse whether this could be another influence on the bad model fit. Nevertheless, factor loadings are comparable. This could be due to the sample size, but also due to the quality of the data. Further research is needed into the BUS-A to clearly state wherein the problem lies. Regarding the second part of the research question, good reliability was found for the BUS-A, but moderate reliability for the separate factors. Thus, the scale itself seems reliable, even though not fitting with the original factorial structure. The study shows that the two assessment instruments might be completely different in their structure.

Again, this finding is in line with the outcomes of the previous research questions. Although not confirmed, the findings implicate that the BUS-A might be accessible, but not in tune with the structure of the original BUS-11. Reliability analyses also show that the scale is reliable, even though the five-factor structure could not have been confirmed. Possibly, barriers could have been removed for disabled people to participate in research (Goegan, 2018), but this did not result in the same assessment instrument. Research does provide some guidelines that can be transferred on how to make research more inclusive, but this does not automatically translate to the fact that the testing instrument will stay the same. This is an important finding, implicating a new research direction.

Finally, running the CFA on the combined data confirmed the suggested factorial model by showing a good fit for the five factors (Borsci et al., 2021b). Therefore, even though

the five-factor structure could not be completely confirmed by the BUS-A itself, combined data validated that structure. Additionally, data can be regarded as reliable, considering the high Cronbach's alpha of over .7, although not completely in tune with the .8 which was aimed at. These findings can be seen in relation to the other satisfaction scales reviewed, such as the SUS, UMUX and CSUQ (Lewis, 2018a). As a usability scale, the BUS can be described as similarly reliable (Lewis, 2018a; Lewis 2018b; Lewis et al., 2013). Considering that it is more specifically tailored to the usability of chatbots, it can be recognized as an important tool to assess the satisfaction of those to then inform redesign recommendations for chatbot providers. This could lead to more specific insights into customer satisfaction than with regular satisfaction scales not tailored to specific systems.

Future Research

Even though the results of the study seem to point towards a clear direction, disparities between the results arise at some point which could be attributed to certain limitations. Two limitations can be identified, which then can be translated into suggestions for future research. One point of improvement of the current study is the small sample size in both phases of it. In the first phase, two participants were recruited for the focus group. Although their opinions were mostly corresponding to each other, further participants could have been more beneficial. Especially when considering that the BUS-A aims at including as many target groups as possible, people with different disabilities would have been helpful in the evaluation process. In contrast to the two participants included in the study, focus groups are recommended to include 6-8 participants (Wilkinson, 1998). Still, focus groups are considered as adding validity to the study and evaluating issues from real-life experiences (Barrett & Kirk, 2000). Furthermore, Barrett and Kirk (2000) report that focus groups are especially useful for designing questionnaires. Therefore, the first limitation of the study could have been a potential strength which was not fully used due to lacking participation in the focus group. The same issue was found in the recruitment of participants for the

experimental group. The second limitation resulting from this is the lacking variety of disabilities of participants in the sample. To test accessibility, a more diverse sample would be preferred.

Lacking participation in the study can be attributed not mainly to the study design, but to a failed cooperation with a partner, who did not deliver results as planned. Initially, collaboration with the Open Minds School was planned on the recruitment of participants with disabilities. Unfortunately, the Open Minds School was not able to recruit any participants for the second part of the study. Afterwards, no further updates were received from the Open Minds School. Therefore, data collection did not proceed as initially planned and a very limited number of responses was collected for the experimental group.

Based on these limitations identified, a suggestion for future research would be to test the BUS-A on a more diverse sample, including people with different disabilities to robustly test its accessibility. Moreover, it is recommended to participate with different organisations for recruiting participants, which could ensure more reliable data collection. In line with that, a further suggestion is to test translations of the BUS-A, which could result in an even more diverse sample of participants. Only if a big enough and diverse enough sample will have tested the BUS-A, clear conclusions regarding its reliability can be drawn. Furthermore, another suggestion aiming at increasing the number of participants is to include another task on another chatbot. Thereby, responses can be more efficiently increased.

Another suggestion for future research is to correlate the BUS-A with other satisfaction scales, such as the UMUX-Lite. Thereby, external validity could be tested, and the accessible version could be validated stronger. Overall, if no difference between BUS-A and BUS-11 is found, this would also validate the original version of the scale. Additionally, this could establish the Chatbot Usability Scale to be more widely accepted. If however differences between the scales are still evident in future research, further studies could also investigate whether there is a relationship between perceived privacy (Factor 4) and time

response of the chatbot (Factor 5). By analysing the nature of the correlation underlying differences between the BUS-A and BUS-11 could be uncovered.

Conclusion

Implementation of chatbots in customer service is increasing, but the problem of lacking accessibility of such systems does not allow people with disabilities to benefit from these. Therefore, the current study aimed at creating a tool usable for everyone to assess chatbots. Thereby, it was important to point out that accessibility is achievable when following specific design recommendations. By showing that there were no differences between people with and without disabilities, the accessibility of the BUS-A has indeed been implicated. Even though differences between the two scales, the BUS-A and the BUS-11 were found, further discussion led to the conclusion that this might have a different cause than the accessible design, which needs to be investigated in future research. Finally, a validation of the BUS itself indicated that it is a reasonable tool to be used in future usability research. Thus, the study could not prove that the new version of the scale (BUS-A) is comparable to the original scale (BUS-11). An accessible scale was created, but it does not fit with the originally implicated model. To enable disabled people to take part in evaluating chatbots and increasing their quality, future studies should work on creating more accessible scales. Thereby, the BUS-A could aid as a reliable tool, if being further validated.

References

- Abbott, S., & Mcconkey, R. (2006). The barriers to social inclusion as perceived by people with intellectual disabilities. *Journal of Intellectual Disabilities*, 10(3), 275–287. <https://doi.org/10.1177/1744629506067618>
- Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*. <https://doi.org/10.1007/s12525-020-00414-7>
- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, 100006. <https://doi.org/10.1016/j.mlwa.2020.100006>
- American Psychiatric Association. (n.d.). *What Is Specific Learning Disorder?*. American Psychiatric Association. Retrieved March 7, 2022, from <https://www.psychiatry.org/patients-families/specific-learning-disorder/what-is-specific-learning-disorder>
- Arias-Durán, S., Sanisaca-Muñoz, J., Bravo-Buri, S., & Robles-Bykbaev, V. (2021). Intervention Platform for Children with Intellectual Disability: Chatbots and IBM Watson Services in the Ecuadorian Context. *Lecture Notes in Networks and Systems*, 486–494. https://doi.org/10.1007/978-3-030-80091-8_57
- Balaji, Divyaa (2019) *Assessing user satisfaction with information chatbots: a preliminary investigation*. Retrieved June 16, 2022, from <https://purl.utwente.nl/essays/79785>
- Barrett, J., & Kirk, S. (2000). Running focus groups with elderly and disabled elderly participants. *Applied Ergonomics*, 31(6), 621–629. [https://doi.org/10.1016/s0003-6870\(00\)00031-4](https://doi.org/10.1016/s0003-6870(00)00031-4)
- Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2021a). 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>
- Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A. & van der Velde, F. (2021b) Confirmatory Factorial Analysis of the Chatbot Usability Scale: A Multilanguage Validation. [Manuscript submitted for publication].
- Bos, M. A. van den (2021) *Testing a scale for perceived usability and user satisfaction in chatbots: Testing the BotScale*. Retrieved June 16, 2022, from <https://purl.utwente.nl/essays/85767>
- Cernat, A., & Liu, M. (2018). Radio buttons in web surveys: Searching for alternatives. *International Journal of Market Research*, 61(3), 266–286. <https://doi.org/10.1177/1470785318813520>
- Color Safe (n.d.). *Accessible Web Color Combinations*. Color Safe. Retrieved April 13, 2022, from <http://colorsafe.co/?msclkid=1ebd229bba5811ec9cdfd58e65a963f3>
- Comrey, A. L., & Lee, H. B. (1992). *A First Course in Factor Analysis* (2nd ed.). Psychology Press. Retrieved June 16, 2022, from <https://www.routledge.com/A-First-Course-in-Factor-Analysis/Comrey-Lee/p/book/9781138965454>
- Davies, D. K., Stock, S. E., King, L., Wehmeyer, M. L., & Shogren, K. A. (2017). An accessible testing, learning and assessment system for people with intellectual disability. *International Journal of Developmental Disabilities*, 63(4), 204–210. <https://doi.org/10.1080/20473869.2017.1294313>
- de Haan, E., Verhoef, P. C., & Wiesel, T. (2015). The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing*, 32(2), 195-206. <https://doi.org/10.1016/j.ijresmar.2015.02.004>

- Easy Read Online. (n.d.). *Questionnaires - Easy-Read-Online Limited*. Easy Read Online. Retrieved March 7, 2022, from <https://www.easy-read-online.co.uk/about-easy-read/questionnaires/>
- Fuchs, L. S., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, 37(6). Retrieved April 29, 2022, from file:///Users/anyaboyko/Downloads/ebk,+foec_v37n6_200502.pdf
- Goegan, L. D. (2018, January 1). *Accessibility in Questionnaire Research*. . . ERA. <https://era.library.ualberta.ca/items/8b8372d9-12b1-474b-b90e-73555300b828>
- Harper, S., & Chen, A. Q. (2011). Web accessibility guidelines. *World Wide Web*, 15(1), 61–88. <https://doi.org/10.1007/s11280-011-0130-8>
- Hartley, S. L., & MacLean, W. E. (2006). A review of the reliability and validity of Likert-type scales for people with intellectual disability. *Journal of Intellectual Disability Research*, 50(11), 813–827. <https://doi.org/10.1111/j.1365-2788.2006.00844.x>
- ISO. (n.d.). *ISO 9241-11:2018(en) Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. <https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en>
- Kerwien Lopez, S.M. (2021) *Confirmatory Factor Analysis of a new Satisfaction Scale for conversational agents and the role of decision-making styles*. Retrieved June 16, 2022, from <https://purl.utwente.nl/essays/86852>
- Kissow, A. M. (2013). Participation in physical activity and the everyday life of people with physical disabilities: a review of the literature. *Scandinavian Journal of Disability Research*, 17(2), 144–166. <https://doi.org/10.1080/15017419.2013.787369>
- Kvale, K., Freddi, E., Hodnebrog, S., Sell, O. A., & Følstad, A. (2021). Understanding the User Experience of Customer Service Chatbots: What Can We Learn from Customer

- Satisfaction Surveys? *Chatbot Research and Design*, 205–218.
https://doi.org/10.1007/978-3-030-68288-0_14
- Lewis, J. R. (2018a). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction*, 34(12), 1148–1156.
<https://doi.org/10.1080/10447318.2017.1418805>
- Lewis, J. R. (2018b). The System Usability Scale: Past, Present, and Future. *International Journal of Human–Computer Interaction*, 34(7), 577–590.
<https://doi.org/10.1080/10447318.2018.1455307>
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
<https://doi.org/10.1145/2470654.2481287>
- Lewthwaite, S. (2014). Web accessibility standards and disability: developing critical perspectives on accessibility. *Disability and Rehabilitation*, 36(16), 1375–1383.
<https://doi.org/10.3109/09638288.2014.938178>
- Mayes, S. D., & Calhoun, S. L. (2007). Learning, Attention, Writing, and Processing Speed in Typical Children and Children with ADHD, Autism, Anxiety, Depression, and Oppositional-Defiant Disorder. *Child Neuropsychology*, 13(6), 469–493.
<https://doi.org/10.1080/09297040601112773>
- Microsoft. (n.d. a). *Make your Word documents accessible to people with disabilities*. Retrieved April 12, 2022, from <https://support.microsoft.com/en-us/office/make-your-word-documents-accessible-to-people-with-disabilities-d9bf3683-87ac-47ea-b91a-78dcacb3c66d>
- Microsoft. (n.d. b). *Microsoft Excel Spreadsheet Software*. Microsoft 365. Retrieved May 16, 2022, from <https://www.microsoft.com/en-us/microsoft-365/excel>
- Moreno, F., Coret, J., Jiménez, E., Márquez, S., & Alcantud, F. (2012). Evaluation of web browsing experience by people with cognitive disability. In *Proceedings of the 13th*

International Conference on Interacción Persona-Ordenador (pp. 1-2).

<https://doi.org/10.1145/2379636.2379673>

Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2020). Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models.

Educational and Psychological Measurement, 81(1), 110–130.

<https://doi.org/10.1177/0013164420926231>

Qualtrics. (n.d.). *Qualtrics XM. The Leading Experience Management Software*. Qualtrics.

Retrieved June 30, 2022, from <https://www.qualtrics.com/uk/>

Rømen, D., & Svanæs, D. (2008). Evaluating web site accessibility: Validating the WAI guidelines through usability testing with disabled users. *Proceedings of the 5th Nordic Conference on Human-Computer Interaction Building Bridges - NordiCHI '08*.

<https://doi.org/10.1145/1463160.1463238>

RPubs. (2015). *RPubs - Reliability analysis with psych package*. RPubs. Retrieved May 22, 2022, from <https://rpubs.com/hauselin/reliabilityanalysis>

RStudio. (n.d.). *RStudio. Open source & professional software for data science teams*.

Retrieved May 16, 2022, from <https://www.rstudio.com/>

Saenz, J., Burgess, W., Gustitis, E., Mena, A., & Sasangohar, F. (2017). The usability analysis of chatbot technologies for internal personnel communications Industrial and Systems Engineering Conference Pittsburgh, Pennsylvania, US. Retrieved June 16, 2022, from [https://www.scopus.com/record/display.uri?eid=2-s2.0-](https://www.scopus.com/record/display.uri?eid=2-s2.0-85031000963&origin=inward&txGid=587fcac3db122ac6f168821e02370db5&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1)

[85031000963&origin=inward&txGid=587fcac3db122ac6f168821e02370db5&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1](https://www.scopus.com/record/display.uri?eid=2-s2.0-85031000963&origin=inward&txGid=587fcac3db122ac6f168821e02370db5&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1)

Sanny, L., Susastra, A. C., Roberts, C., & Yusramdaleni, R. (2020). The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia. *Management Science Letters*, 1225–1232.

<https://doi.org/10.5267/j.msl.2019.11.036>

- Schrepp, M., Pérez Cota, M., Gonçalves, R., Hinderks, A., & Thomaschewski, J. (2016). Adaption of user experience questionnaires for different user groups. *Universal Access in the Information Society*, 16(3), 629–640. <https://doi.org/10.1007/s10209-016-0485-9>
- Si, S. (2020, December 1). *What Does Having an “Accessible” Survey Mean? – Conversion Rate*. . . Qeryz. Retrieved April 12, 2022, from <https://qeryz.com/blog/accessible-survey/>
- Sik Lanyi, C. (2017). Choosing effective colours for websites. *Colour Design*, 619–640. <https://doi.org/10.1016/b978-0-08-101270-3.00026-6>
- Silderhuis, I. (2020) *Validity and Reliability of the User Satisfaction with Information Chatbots Scale (USIC)*. Retrieved June 16, 2022, from <https://purl.utwente.nl/essays/83495>
- Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The Search for “Optimal” Cutoff Properties: Fit Index Criteria in Structural Equation Modeling. *The Journal of Experimental Education*, 74(3), 267–288. <https://doi.org/10.3200/jexe.74.3.267-288>
- Spina, C. (2019). WCAG 2.1 and the Current State of Web Accessibility in Libraries. *Weave: Journal of Library User Experience*, 2(2). <https://doi.org/10.3998/weave.12535642.0002.202>
- Spire Digital. (2019). *6 Principles of Accessible Design - UX Planet*. Medium. Retrieved April 23, 2022, from <https://uxplanet.org/6-principles-of-accessible-design-79beceeaaffb>
- Torres, C., Franklin, W., & Martins, L. (2018). Accessibility in Chatbots: The State of the Art in Favor of Users with Visual Impairment. *Advances in Usability, User Experience and Assistive Technology*, 623–635. https://doi.org/10.1007/978-3-319-94947-5_63

Tsatsou, P. (2020). Is digital inclusion fighting disability stigma? Opportunities, barriers, and recommendations. *Disability & Society*, 36(5), 702–729.

<https://doi.org/10.1080/09687599.2020.1749563>

University of Washington. (n.d.). *Creating Accessible Documents*. University of Washington.

Retrieved April 12, 2022, from <https://www.washington.edu/accessibility/documents/>

Utwente Sona Systems. (n.d.). *Test Subject Pool BMS*. Sona Systems. Retrieved June 30,

2022, from <https://utwente.sona-systems.com/Default.aspx?ReturnUrl=%2f>

Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Canello, H. (2017). *Here's What I Can*

Do: Chatbots' Strategies to Convey Their Features to Users The XVI Brazilian

Symposium on Human Factors in Computing Systems, Joinville, Brazil.

<https://doi.org/10.1145/3160504.3160544>

Vanderheiden, G. C., Jordan, J. B., & Lazar, J. (2021). DESIGN FOR PEOPLE

EXPERIENCING FUNCTIONAL LIMITATIONS. *HANDBOOK OF HUMAN*

FACTORS AND ERGONOMICS, 1216–1248.

<https://doi.org/10.1002/9781119636113.ch47>

Vereenooghe, L. (2021). Participation of People With Disabilities in Web-Based Research.

Zeitschrift Für Psychologie, 229(4), 257–259. <https://doi.org/10.1027/2151->

[2604/a000472](https://doi.org/10.1027/2151-2604/a000472)

W3C. (2018, June 5). *Web Content Accessibility Guidelines (WCAG) 2.1*. W3C.

<https://www.w3.org/TR/WCAG21/>

Walmsley, J. (2001). Normalisation, Emancipatory Research and Inclusive Research in

Learning Disability. *Disability & Society*, 16(2), 187–205.

<https://doi.org/10.1080/09687590120035807>

Wilkinson, S. (1998). Focus group methodology: a review. *International Journal of Social*

Research Methodology, 1(3), 181–203.

<https://doi.org/10.1080/13645579.1998.10846874>

World Health Organization. (n.d.). *Disability and health*. <https://www.who.int/news-room/fact-sheets/detail/disability-and-health>

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12), 2141–2155.

<https://doi.org/10.1080/00949655.2010.520163>

Zumstein, D., & Hundertmark, S. (2017). CHATBOTS--AN INTERACTIVE TECHNOLOGY FOR PERSONALIZED COMMUNICATION, TRANSACTIONS AND SERVICES. *IADIS International Journal on WWW/Internet*, 15(1).

[https://www.researchgate.net/profile/Darius-](https://www.researchgate.net/profile/Darius-Zumstein/publication/322855718_Chatbots_-_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_Services/links/5a72ecde458515512076b406/Chatbots-An-Interactive-Technology-for-Personalized-Communication-Transactions-and-Services.pdf)

[Zumstein/publication/322855718_Chatbots_-](https://www.researchgate.net/profile/Darius-Zumstein/publication/322855718_Chatbots_-_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_Services/links/5a72ecde458515512076b406/Chatbots-An-Interactive-Technology-for-Personalized-Communication-Transactions-and-Services.pdf)

[_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_S](https://www.researchgate.net/profile/Darius-Zumstein/publication/322855718_Chatbots_-_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_Services/links/5a72ecde458515512076b406/Chatbots-An-Interactive-Technology-for-Personalized-Communication-Transactions-and-Services.pdf)

[ervices/links/5a72ecde458515512076b406/Chatbots-An-Interactive-Technology-for-](https://www.researchgate.net/profile/Darius-Zumstein/publication/322855718_Chatbots_-_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_Services/links/5a72ecde458515512076b406/Chatbots-An-Interactive-Technology-for-Personalized-Communication-Transactions-and-Services.pdf)

[Personalized-Communication-Transactions-and-Services.pdf](https://www.researchgate.net/profile/Darius-Zumstein/publication/322855718_Chatbots_-_An_Interactive_Technology_for_Personalized_Communication_Transactions_and_Services/links/5a72ecde458515512076b406/Chatbots-An-Interactive-Technology-for-Personalized-Communication-Transactions-and-Services.pdf)

Appendix A - Informed Consent Focus Group

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 B - Focus Group Protocol

Focus Group Protocol

Scheduled on Monday, Apr 4, 2022

Interviewer: I

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 C - Task Instructions Focus Group

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 - Prototype BUS-A 1

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.

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

1

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

2






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.






Two statements will be presented on one page, except for the first page. On this page, you will only find one item.

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

3

2. It was easy to find the chatbot. 🔍

<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

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

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

5

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

<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

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

<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

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

1

STRONGLY
DISAGREE

2



DISAGREE

3

NEITHER
DISAGREE NOR
AGREE

4



AGREE

5

STRONGLY
AGREE

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

7

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!

8

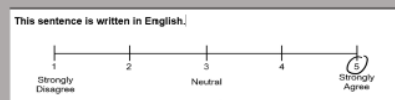
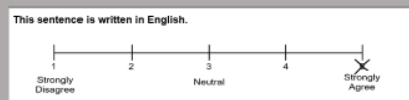
Appendix E - Prototype BUS-A 2

Chatbot Usability Questionnaire

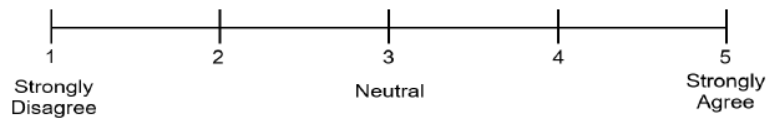
Section one instructions:

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.

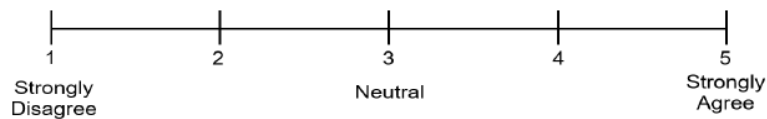
This is an exemplary sentence with an answer:



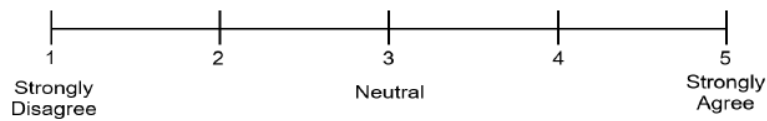
1. I am familiar with chatbots or other conversational agents.



2. I know how chatbots work.



3. I am confident in using chatbots.

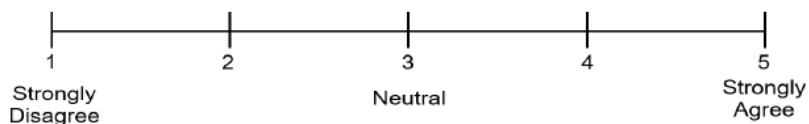


If you have performed the task provided to you fill in the remaining part of the questionnaire. If not, please perform the task and then come back to the questionnaire.

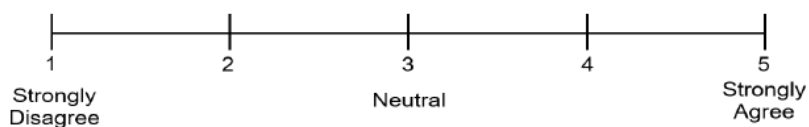
Section two instructions:

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.

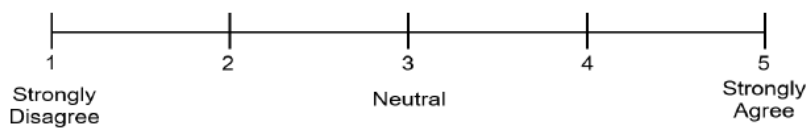
1. The chatbot function was easily detectable.



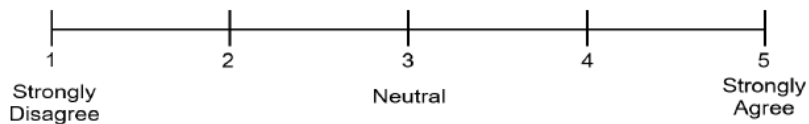
2. It was easy to find the chatbot.



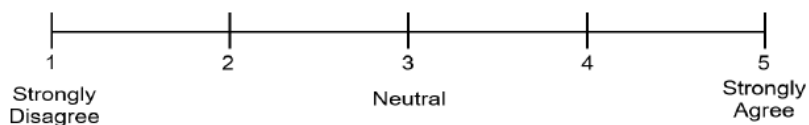
3. Communicating with the chatbot was clear.



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



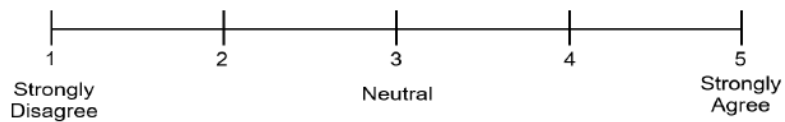
5. The chatbot's responses were easy to understand.



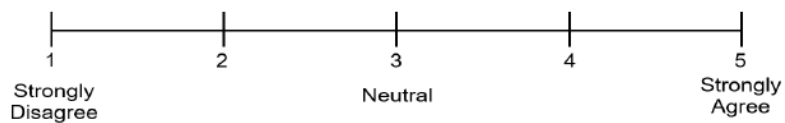
Continue to next page

Continuation of previous page

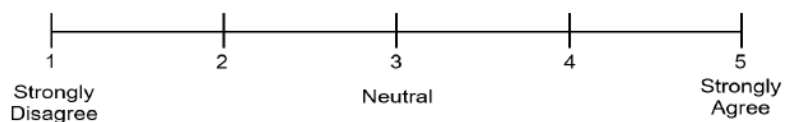
6. I find that the chatbot understands what I want and helps me achieve my goal.



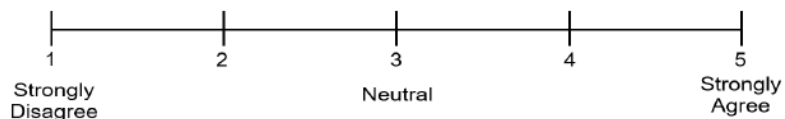
7. The chatbot gives me the appropriate amount of information.



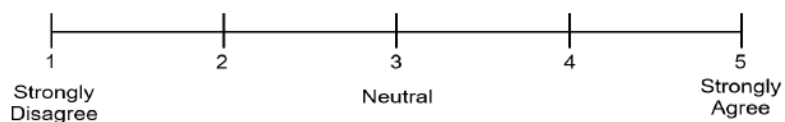
8. The chatbot gives me the information I need.



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



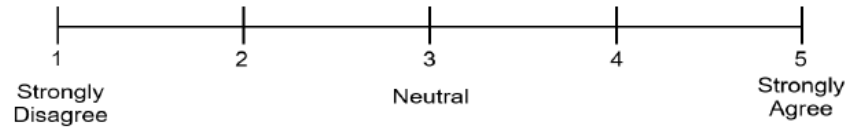
10. I believe the chatbot informs me of any possible privacy issues.



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 - Transcription of the Focus Group

Open Mind School
Chatbot Accessibility Focus Group
Interview 1

Interviewer	E
Communication Partner	C
Participant 1	P1
Participant 2	P2

P1 [16:35 in recording]

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

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

P1: Easy

E: 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.

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

P1: Yes.

E: Did you like the answering options?

P1: No.

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

P1: No.

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

C: Would you want it bigger or smaller?

P1: Bigger.

E: 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.

E: Would you change the scale?

P1: No.

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

P1: Too long.

E: 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]

E: 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]

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

P2: Yes.

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

P2: Yes.

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

P2: That one and this one.

E: Were there any more?

P2: No.

E: Just the first two?

P2: Yeah.

E: 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.

E: Did you like the layout?

P2: Yes.

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

P2: Everything.

E: Did you say everything?

P2: Yeah.

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

P2: No.

E: Okay. How was the font and size?

P2: Bad.

E: Bad? How would you change it?

P2: I don't know.

E: Would you make the text smaller or bigger?

P2: Bigger.

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

P2: Yes.

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

P2: Yes.

E: Can you tell me about those?

P2: Yes.

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

P2: Yes.

E: Did you like or dislike it?

P2: Liked it.

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

P2: Liked.

E: 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]

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

P2: Because I like it.

E: 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 - Final Version – BUS-A

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 strongly you agree with the statements. The choices range from “strongly disagree” to “strongly agree”.

To answer, select the choice that you agree with the most.

N

Demographics

Before actually conducting the questionnaire, please 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

N

2. Do you consider yourself to have a disability that can affect your experience with the chatbots (e.g. developmental disability, learning disability, mental health or emotional disability, unseen disability, physical disability, sensory disability, etc.)?

Please select one answer.

Yes

No

3

N 

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 can 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

N 

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

- 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

N

5. What was your sex as assigned at birth?

Please select one answer. ☒

Male

Female

6. Are you already familiar with chatbots?

Please select one answer. ☒

Yes

No

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

6

N

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, select the choice that you agree with the most. ☒

Please choose **only one** answer.

One statement will be presented per page.

7

N

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

8

N

1. The chatbot function was easily detectable. 🔍

1



STRONGLY
DISAGREE

2



DISAGREE

3



NEITHER
DISAGREE
NOR AGREE

4



AGREE

5






STRONGLY
AGREE

9

N



2. It was easy to find the chatbot. 🔍

<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

N




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

11

N






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

12


N




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

13

N






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

<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

14


N






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

<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

15

N


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

16


N

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

17

N

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

18

N

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!

19

Appendix H - RStudio Code

```
#libraries  
  
library(readxl)  
  
library(readxl)  
  
library(foreign)  
  
library(lavaan)  
  
library(lavaanPlot)  
  
library(dplyr)  
  
library(haven)  
  
library(ggpubr)  
  
library(knitr)  
  
library(semPlot)  
  
library(MVN)  
  
library(tidyr)  
  
library(tidyverse)  
  
library(WriteXLS)  
  
library(ltm)  
  
library(outliers)  
  
library(EnvStats)  
  
library (polycor)  
  
library (ggpubr)  
  
library(rstatix)  
  
library (psych)  
  
library(ggplot2)  
  
library(dplyr)  
  
library(broom)
```

```
library(ggpubr)

library(car)

#importing the data

cont <- read_excel("Revised_Control.xlsx")

exp <- read_excel("Experimental.xlsx")

all <- read_excel("Revised.xlsx")

view(cont) # Item_1

view(exp) #i1 etc

view(all) # Item_1

#####turn off scientific notation

options(scipen = 999)

# turning scores into percentages

q1 <- cont$Item_1

q2 <- cont$Item_2

q3 <- cont$Item_3

q4 <- cont$Item_4

q5 <- cont$Item_5

q6 <- cont$Item_6

q7 <- cont$Item_7

q8 <- cont$Item_8
```

```
q9 <- cont$Item_9
q10 <- cont$Item_10
q11 <- cont$Item_11

cont$sum <- q1+q2+q3+q4+q5+q6+q7+q8+q9+q10+q11
cont$perc <- as.numeric(cont$sum) /55
cont$perc <- paste(round(100*cont$perc, 2),"% ", sep="")

percent_vec = paste(1:100, "% ", sep = "")
cont$perc <- paste(as.numeric(sub("% ", "", cont$perc)))

#score per item
cont$I1 <- as.numeric(cont$Item_1) /5
cont$I1 <- paste(round(100*cont$I1, 2),"% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
cont$I1 <- paste(as.numeric(sub("% ", "", cont$I1)))

cont$I2 <- as.numeric(cont$Item_2) /5
cont$I2 <- paste(round(100*cont$I2, 2),"% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
cont$I2 <- paste(as.numeric(sub("% ", "", cont$I2)))

cont$I3 <- as.numeric(cont$Item_3) /5
cont$I3 <- paste(round(100*cont$I3, 2),"% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
cont$I3 <- paste(as.numeric(sub("% ", "", cont$I3)))
```

```
cont$I4 <- as.numeric(cont$Item_4) /5
cont$I4 <- paste(round(100*cont$I4, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$I4 <- paste(as.numeric(sub("%", "", cont$I4)))

cont$I5 <- as.numeric(cont$Item_5) /5
cont$I5 <- paste(round(100*cont$I5, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$I5 <- paste(as.numeric(sub("%", "", cont$I5)))

cont$I6 <- as.numeric(cont$Item_6) /5
cont$I6 <- paste(round(100*cont$I6, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$I6 <- paste(as.numeric(sub("%", "", cont$I6)))

cont$I7 <- as.numeric(cont$Item_7) /5
cont$I7 <- paste(round(100*cont$I7, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$I7 <- paste(as.numeric(sub("%", "", cont$I7)))

cont$I8 <- as.numeric(cont$Item_8) /5
cont$I8 <- paste(round(100*cont$I8, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$I8 <- paste(as.numeric(sub("%", "", cont$I8)))
```

```
cont$I9 <- as.numeric(cont$Item_9) /5
cont$I9 <- paste(round(100*cont$I9, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
cont$I9 <- paste(as.numeric(sub("%", "", cont$I9)))

cont$I10 <- as.numeric(cont$Item_10) /5
cont$I10 <- paste(round(100*cont$I10, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
cont$I10 <- paste(as.numeric(sub("%", "", cont$I10)))

cont$I11 <- as.numeric(cont$Item_11) /5
cont$I11 <- paste(round(100*cont$I11, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
cont$I11 <- paste(as.numeric(sub("%", "", cont$I11)))

view(cont)

#experimental
qq1<- exp$i2`
qq2<- exp$i2`
qq3<- exp$i3`
qq4<- exp$i4`
qq5<- exp$i5`
qq6<- exp$i6`
qq7<- exp$i7`
qq8<- exp$i8`
```



```
qq9<- exp$i9`
```

```
qq10<- exp$i10`
```

```
qq11<- exp$i11`
```

```
exp$sum <- qq1+qq2+qq3+qq4+qq5+qq6+qq7+qq8+qq9+qq10+qq11
```

```
exp$perc <- as.numeric(exp$sum) /55
```

```
exp$perc <- paste(round(100*exp$perc, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$perc <- paste(as.numeric(sub("%", "", exp$perc)))
```

```
view(exp)
```

```
####factors for experimental
```

```
exp$f1 <- qq1+qq2
```

```
exp$f1 <- as.numeric(exp$f1) /10
```

```
exp$f2 <- qq3+qq4+qq5
```

```
exp$f2 <- as.numeric(exp$f2) /15
```

```
exp$f3 <- qq6+qq7+qq8+qq9
```

```
exp$f3 <- as.numeric(exp$f3) /20
```

```
exp$f4 <- qq10
```

```
exp$f4 <- as.numeric(exp$f4) /5
```

```
exp$f5 <- qq11
```

```
exp$f5 <- as.numeric(exp$f5) /5
```

```
exp$f1 <- paste(round(100*exp$f1, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$f1 <- paste(as.numeric(sub("%", "", exp$f1)))
```

```
exp$f2 <- paste(round(100*exp$f2, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$f2 <- paste(as.numeric(sub("%", "", exp$f2)))
```

```
exp$f3 <- paste(round(100*exp$f3, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$f3 <- paste(as.numeric(sub("%", "", exp$f3)))
```

```
exp$f4 <- paste(round(100*exp$f4, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$f4 <- paste(as.numeric(sub("%", "", exp$f4)))
```

```
exp$f5 <- paste(round(100*exp$f5, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$f5 <- paste(as.numeric(sub("%", "", exp$f5)))
```

```
view(exp)
```

```
#items for experimental
```

```
exp$I1 <- as.numeric(exp$i1) /5
```

```
exp$I1 <- paste(round(100*exp$I1, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "% ", sep = "")
exp$I1 <- paste(as.numeric(sub("% ", "", exp$I1)))

exp$I2 <- as.numeric(exp$i2) /5
exp$I2 <- paste(round(100*exp$I2, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
exp$I2 <- paste(as.numeric(sub("% ", "", exp$I2)))

exp$I3 <- as.numeric(exp$i3) /5
exp$I3 <- paste(round(100*exp$I3, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
exp$I3 <- paste(as.numeric(sub("% ", "", exp$I3)))

exp$I4 <- as.numeric(exp$i4) /5
exp$I4 <- paste(round(100*exp$I4, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
exp$I4 <- paste(as.numeric(sub("% ", "", exp$I4)))

exp$I5 <- as.numeric(exp$i5) /5
exp$I5 <- paste(round(100*exp$I5, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
exp$I5 <- paste(as.numeric(sub("% ", "", exp$I5)))

exp$I6 <- as.numeric(exp$i6) /5
exp$I6 <- paste(round(100*exp$I6, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
```

```
exp$I6 <- paste(as.numeric(sub("%", "", exp$I6)))
```

```
exp$I7 <- as.numeric(exp$i7) /5
```

```
exp$I7 <- paste(round(100*exp$I7, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$I7 <- paste(as.numeric(sub("%", "", exp$I7)))
```

```
exp$I8 <- as.numeric(exp$i8) /5
```

```
exp$I8 <- paste(round(100*exp$I8, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$I8 <- paste(as.numeric(sub("%", "", exp$I8)))
```

```
exp$I9 <- as.numeric(exp$i9) /5
```

```
exp$I9 <- paste(round(100*exp$I9, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$I9 <- paste(as.numeric(sub("%", "", exp$I9)))
```

```
exp$I10 <- as.numeric(exp$i10) /5
```

```
exp$I10 <- paste(round(100*exp$I10, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$I10 <- paste(as.numeric(sub("%", "", exp$I10)))
```

```
exp$I11 <- as.numeric(exp$i11) /5
```

```
exp$I11 <- paste(round(100*exp$I11, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
exp$I11 <- paste(as.numeric(sub("%", "", exp$I11)))
```

```

view(exp)

#factors as percentages

cont$f1 <- q1+q2
cont$f1 <- as.numeric(cont$f1) /10

cont$f2 <- q3+q4+q5
cont$f2 <- as.numeric(cont$f2) /15

cont$f3 <- q6+q7+q8+q9
cont$f3 <- as.numeric(cont$f3) /20

cont$f4 <- q10
cont$f4 <- as.numeric(cont$f4) /5

cont$f5 <- q11
cont$f5 <- as.numeric(cont$f5) /5

cont$f1 <- paste(round(100*cont$f1, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$f1 <- paste(as.numeric(sub("%", "", cont$f1)))

cont$f2 <- paste(round(100*cont$f2, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$f2 <- paste(as.numeric(sub("%", "", cont$f2)))

cont$f3 <- paste(round(100*cont$f3, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
cont$f3 <- paste(as.numeric(sub("%", "", cont$f3)))

```

```
cont$f4 <- paste(round(100*cont$f4, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
cont$f4 <- paste(as.numeric(sub("%", "", cont$f4)))
```

```
cont$f5 <- paste(round(100*cont$f5, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
cont$f5 <- paste(as.numeric(sub("%", "", cont$f5)))
```

```
view(cont)
```

```
#####transforming data set all qa <- all
```

```
qa1 <- all$Item_1
```

```
qa2 <- all$Item_2
```

```
qa3 <- all$Item_3
```

```
qa4 <- all$Item_4
```

```
qa5 <- all$Item_5
```

```
qa6 <- all$Item_6
```

```
qa7 <- all$Item_7
```

```
qa8 <- all$Item_8
```

```
qa9 <- all$Item_9
```

```
qa10 <- all$Item_10
```

```
qa11 <- all$Item_11
```

```
all$sum <- qa1+qa2+qa3+qa4+qa5+qa6+qa7+qa8+qa9+qa10+qa11
```

```
all$perc <- as.numeric(all$sum) /55
all$perc <- paste(round(100*all$perc, 2),"%", sep="")

percent_vec = paste(1:100, "%", sep = "")
all$perc <- paste(as.numeric(sub("%", "", all$perc)))

view(all)

#score per item
all$I1 <- as.numeric(all$Item_1) /5
all$I1 <- paste(round(100*all$I1, 2),"%", sep="")
percent_vec = paste(1:100, "%", sep = "")
all$I1 <- paste(as.numeric(sub("%", "", all$I1)))

all$I2 <- as.numeric(all$Item_2) /5
all$I2 <- paste(round(100*all$I2, 2),"%", sep="")
percent_vec = paste(1:100, "%", sep = "")
all$I2 <- paste(as.numeric(sub("%", "", all$I2)))

all$I3 <- as.numeric(all$Item_3) /5
all$I3 <- paste(round(100*all$I3, 2),"%", sep="")
percent_vec = paste(1:100, "%", sep = "")
all$I3 <- paste(as.numeric(sub("%", "", all$I3)))

all$I4 <- as.numeric(all$Item_4) /5
all$I4 <- paste(round(100*all$I4, 2),"%", sep="")
```

```
percent_vec = paste(1:100, "% ", sep = "")
all$I4 <- paste(as.numeric(sub("% ", "", all$I4)))

all$I5 <- as.numeric(all$Item_5) /5
all$I5 <- paste(round(100*all$I5, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
all$I5 <- paste(as.numeric(sub("% ", "", all$I5)))

all$I6 <- as.numeric(all$Item_6) /5
all$I6 <- paste(round(100*all$I6, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
all$I6 <- paste(as.numeric(sub("% ", "", all$I6)))

all$I7 <- as.numeric(all$Item_7) /5
all$I7 <- paste(round(100*all$I7, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
all$I7 <- paste(as.numeric(sub("% ", "", all$I7)))

all$I8 <- as.numeric(all$Item_8) /5
all$I8 <- paste(round(100*all$I8, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
all$I8 <- paste(as.numeric(sub("% ", "", all$I8)))

all$I9 <- as.numeric(all$Item_9) /5
all$I9 <- paste(round(100*all$I9, 2), "% ", sep="")
percent_vec = paste(1:100, "% ", sep = "")
```



```
all$I9 <- paste(as.numeric(sub("%", "", all$I9)))

all$I10 <- as.numeric(all$Item_10) /5
all$I10 <- paste(round(100*all$I10, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
all$I10 <- paste(as.numeric(sub("%", "", all$I10)))

all$I11 <- as.numeric(all$Item_11) /5
all$I11 <- paste(round(100*all$I11, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
all$I11 <- paste(as.numeric(sub("%", "", all$I11)))

#factors as percentages
all$f1 <- qa1+qa2
all$f1 <- as.numeric(all$f1) /10
all$f2 <- qa3+qa4+qa5
all$f2 <- as.numeric(all$f2) /15
all$f3 <- qa6+qa7+qa8+qa9
all$f3 <- as.numeric(all$f3) /20
all$f4 <- qa10
all$f4 <- as.numeric(all$f4) /5
all$f5 <- qa11
all$f5 <- as.numeric(all$f5) /5

all$f1 <- paste(round(100*all$f1, 2), "%", sep="")
percent_vec = paste(1:100, "%", sep = "")
```

```
all$f1 <- paste(as.numeric(sub("%", "", all$f1)))
```

```
all$f2 <- paste(round(100*all$f2, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
all$f2 <- paste(as.numeric(sub("%", "", all$f2)))
```

```
all$f3 <- paste(round(100*all$f3, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
all$f3 <- paste(as.numeric(sub("%", "", all$f3)))
```

```
all$f4 <- paste(round(100*all$f4, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
all$f4 <- paste(as.numeric(sub("%", "", all$f4)))
```

```
all$f5 <- paste(round(100*all$f5, 2), "%", sep="")
```

```
percent_vec = paste(1:100, "%", sep = "")
```

```
all$f5 <- paste(as.numeric(sub("%", "", all$f5)))
```

```
view(all)
```

```
#####transforming the new variables into numeric
```

```
all$I1 <- as.numeric(all$I1)
```

```
class(all$I1)
```

```
all$I2 <- as.numeric(all$I2)
```

```
class(all$I2)
```

```
all$I3 <- as.numeric(all$I3)
```

```
class(all$I3)
```

```
all$I4 <- as.numeric(all$I4)
```

```
class(all$I4)
```

```
all$I5 <- as.numeric(all$I5)
```

```
class(all$I5)
```

```
all$I6 <- as.numeric(all$I6)
```

```
class(all$I6)
```

```
all$I7 <- as.numeric(all$I7)
```

```
class(all$I7)
```

```
all$I8 <- as.numeric(all$I8)
```

```
class(all$I8)
```

```
all$I9 <- as.numeric(all$I9)
```

```
class(all$I9)
```

```
all$I10 <- as.numeric(all$I10)
```

```
class(all$I10)
```

```
all$I11 <- as.numeric(all$I11)
```

```
class(all$I11)
```

```
all$perc <- as.numeric(all$perc)
```

```
class(all$perc)
```

```
all$f1 <- as.numeric(all$f1)
```

```
class(all$f1)
```

```
all$f2 <- as.numeric(all$f2)
```

```
class(all$f2)
```

```
all$f3 <- as.numeric(all$f3)
```

```
class(all$f3)
```

```
all$f4 <- as.numeric(all$f4)
```

```
class(all$f4)
```

```
all$f5 <- as.numeric(all$f5)
```

```
class(all$f5)
```

```
summary(all)
```

```
###for cont dataset
```

```
cont$I1 <- as.numeric(cont$I1)
```

```
class(cont$I1)
```

```
cont$I2 <- as.numeric(cont$I2)
```

```
class(cont$I2)
```

```
cont$I3 <- as.numeric(cont$I3)
```

```
class(cont$I3)
```

```
cont$I4 <- as.numeric(cont$I4)
```

```
class(cont$I4)
```

```
cont$I5 <- as.numeric(cont$I5)
```

```
class(cont$I5)
```

```
cont$I6 <- as.numeric(cont$I6)
```

```
class(cont$I6)
```

```
cont$I7 <- as.numeric(cont$I7)
```

```
class(cont$I7)
```

```
cont$I8 <- as.numeric(cont$I8)
```

```
class(cont$I8)
```

```
cont$I9 <- as.numeric(cont$I9)
```

```
class(cont$I9)
```

```
cont$I10 <- as.numeric(cont$I10)
```

```
class(cont$I10)
```

```
cont$I11 <- as.numeric(cont$I11)
```

```
class(cont$I11)
```

```
cont$perc <- as.numeric(cont$perc)
```

```
class(cont$perc)
```

```
cont$f1 <- as.numeric(cont$f1)
```

```
class(cont$f1)
```

```
cont$f2 <- as.numeric(cont$f2)
```

```
class(cont$f2)
```

```
cont$f3 <- as.numeric(cont$f3)
```

```
class(cont$f3)
```

```
cont$f4 <- as.numeric(cont$f4)
```

```
class(cont$f4)
```

```
cont$f5 <- as.numeric(cont$f5)
```

```
class(cont$f5)
```

```
#### for exp dataset
```

```
exp$I1 <- as.numeric(exp$I1)
```

```
class(exp$I1)
```

```
exp$I2 <- as.numeric(exp$I2)
```

```
class(exp$I2)
```

```
exp$I3 <- as.numeric(exp$I3)
```

```
class(exp$I3)
```

```
exp$I4 <- as.numeric(exp$I4)
```

```
class(exp$I4)
```

```
exp$I5 <- as.numeric(exp$I5)
```

```
class(exp$I5)
```

```
exp$I6 <- as.numeric(exp$I6)
```

```
class(exp$I6)
```

```
exp$I7 <- as.numeric(exp$I7)
```

```
class(exp$I7)
```

```
exp$I8 <- as.numeric(exp$I8)
```

```
class(exp$I8)
```

```
exp$I9 <- as.numeric(exp$I9)
```

```
class(exp$I9)
```

```
exp$I10 <- as.numeric(exp$I10)
```

```
class(exp$I10)
```

```
exp$I11 <- as.numeric(exp$I11)
```

```
class(exp$I11)
```

```
exp$perc <- as.numeric(exp$perc)
```

```
class(exp$perc)
```

```
exp$f1 <- as.numeric(exp$f1)
```

```
class(exp$f1)
```

```
exp$f2 <- as.numeric(exp$f2)
```

```
class(exp$f2)
```

```
exp$f3 <- as.numeric(exp$f3)
```

```
class(exp$f3)
```

```
exp$f4 <- as.numeric(exp$f4)
```

```
class(exp$f4)
```

```
exp$f5 <- as.numeric(exp$f5)
```

```
class(exp$f5)
```

```
#####descriptives
```



```
summary(all)
```

```
sd(all$I1, na.rm = TRUE)
```

```
sd(all$I2, na.rm = TRUE)
```

```
sd(all$I3, na.rm = TRUE)
```

```
sd(all$I4, na.rm = TRUE)
```

```
sd(all$I5, na.rm = TRUE)
```

```
sd(all$I6, na.rm = TRUE)
```

```
sd(all$I7, na.rm = TRUE)
```

```
sd(all$I8, na.rm = TRUE)
```

```
sd(all$I9, na.rm = TRUE)
```

```
sd(all$I10, na.rm = TRUE)
```

```
sd(all$I11, na.rm = TRUE)
```

```
sd(all$perc, na.rm = TRUE)
```

```
####Normality testing
```

```
#QQ Plots still missing
```

```
#Shapiro
```

```
shapiro.test(all$I1)
```

```
shapiro.test(all$I2)
```

```
shapiro.test(all$I3)
```

```
shapiro.test(all$I4)
```

```
shapiro.test(all$I5)
```

```
shapiro.test(all$I6)
```

```
shapiro.test(all$I7)
```

```
shapiro.test(all$I8)
```

```
shapiro.test(all$I9)
```

```
shapiro.test(all$I10)
```

```
shapiro.test(all$I11)
```

```
shapiro.test(all$perc)
```

```
#####t-test
```

```
mean.exp <- exp$perc
```

```
mean.cont <- cont$perc
```

```
t.test(mean.exp, mean.cont, alternative = "two.sided", var.equal = FALSE)
```

```
##for factors
```

```
cf1 <- rowMeans(cont[,c("I1", "I2")], na.rm=TRUE)
```

```
cf2 <- rowMeans(cont[,c("I3", "I4", "I5")], na.rm=TRUE)
```

```
cf3 <- rowMeans(cont[,c("I6", "I7", "I8", "I9")], na.rm=TRUE)
```

```
cf4 <- rowMeans(cont[,c("I10")], na.rm=TRUE)
```

```
cf5 <- rowMeans(cont[,c("I11")], na.rm=TRUE)
```

```
ef1 <- rowMeans(exp[,c("I1", "I2")], na.rm=TRUE)
```

```
ef2 <- rowMeans(exp[,c("I3", "I4", "I5")], na.rm=TRUE)
```

```
ef3 <- rowMeans(exp[,c("I6", "I7", "I8", "I9")], na.rm=TRUE)
```

```
ef4 <- rowMeans(exp[,c("I10")], na.rm=TRUE)
```

```
ef5 <- rowMeans(exp[,c("I11")], na.rm=TRUE)
```

```
t.test(cf1, ef1, alternative = "two.sided", var.equal = FALSE)
```

```
t.test(cf2, ef2, alternative = "two.sided", var.equal = TRUE)
```

```
t.test(cf3, ef3, alternative = "two.sided", var.equal = FALSE)
```

```
t.test(cf4, ef4, alternative = "two.sided", var.equal = FALSE)
```

```
t.test(cf5, ef5, alternative = "two.sided", var.equal = FALSE)
```

```
#####regression analysis
```

```
#need new variable mean satisfaction score for each person
```

```
alpha(all[,c("I1", "I2", "I3", "I4", "I5", "I6", "I7", "I8", "I9", "I10", "I11")])
```

```
all$mean <- rowMeans(all[,c("I1", "I2", "I3", "I4", "I5", "I6", "I7", "I8", "I9", "I10", "I11")],
na.rm=TRUE)
```

```
#linear regression
```

```
group.mean.lm <- lm(mean ~ Group, data = all)
```

```
summary(group.mean.lm)
```

```
#####CFA
```

```
#####CFA for cont
```

```
model1 <- 'fac1 =~ I1+I2
```

```
      fac2 =~ I3 + I4 + I5
```

```
      fac3 =~ I6 + I7 + I8 + I9
```

```
      fac4 =~ I10
```

```
      fac5 =~ I11'
```

```
#Model that we want to confirm #has been run
```

```
#CFA Model Cont
```

```
#deleting columns not needed for the analysis
```

```
cont$chatbot <- NULL
```

```
cont$Item_1 <- NULL
```

```
cont$Item_2 <- NULL
```

```
cont$Item_3 <- NULL
```

```
cont$Item_4 <- NULL
```

```
cont$Item_5 <- NULL
```

```
cont$Item_6 <- NULL
```

```
cont$Item_7 <- NULL
```

```
cont$Item_8 <- NULL
```

```
cont$Item_9 <- NULL
```

```
cont$Item_10 <- NULL
```

```
cont$Item_11 <- NULL
```

```
cont$sum <- NULL
```

```
cont$perc <- NULL
```

```
cont$f1 <- NULL
```

```
cont$f2 <- NULL
```

```
cont$f3 <- NULL
```

```
cont$f4 <- NULL
```

```
cont$f5 <- NULL
```

```
view(cont)
```

```
fit <- cfa(model1, data = cont, estimator="MLR", mimic="Mplus")
```

```
summary(fit, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE)
```

```
####graphical Model:
```

```
semPaths(fit,whatLabels="std",edge.label.cex=1, style = "lisrel", residScale=8, layout  
="tree3", theme = "colorblind", rotation= 2, what="std", nChartNodes = 0, curvePivot=  
TRUE, sizeMan = 4, sizeLat = 10)
```

```
#CFA Model exp
```

```
exp$Participant <- NULL
```

```
exp$Progress <- NULL
```

```
exp$Finished <- NULL
```

```
exp$Requirements_met <- NULL
```

```
exp$Consent <- NULL
```

```
exp$disability_general <- NULL
```

```
exp$disability_specific <- NULL
```

```
exp$Age_E <- NULL
```

```
exp$Gender_E <- NULL
exp$SaB_E <- NULL
exp$Chatbot_Fam_E <- NULL
exp$feedback <- NULL
exp$i1 <- NULL
exp$i2 <- NULL
exp$i3 <- NULL
exp$i4 <- NULL
exp$i5 <- NULL
exp$i6 <- NULL
exp$i7 <- NULL
exp$i8 <- NULL
exp$i9 <- NULL
exp$i10 <- NULL
exp$i11 <- NULL
exp$sum <- NULL
exp$perc <- NULL
exp$f1 <- NULL
exp$f2 <- NULL
exp$f3 <- NULL
exp$f4 <- NULL
exp$f5 <- NULL
view(exp)

fit <- cfa(model1, data = exp, estimator="MLR", mimic="Mplus")
summary(fit, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE)
```

```
####graphical Model:
```

```
semPaths(fit,whatLabels="std",edge.label.cex=1, style = "lisrel", residScale=8, layout  
="tree3", theme = "colorblind", rotation= 2, what="std", nChartNodes = 0, curvePivot=  
TRUE, sizeMan = 4, sizeLat = 10)
```

```
#####CFA with all data
```

```
all$chatbot <- NULL
```

```
all$Item_1 <- NULL
```

```
all$Item_2 <- NULL
```

```
all$Item_3 <- NULL
```

```
all$Item_4 <- NULL
```

```
all$Item_5 <- NULL
```

```
all$Item_6 <- NULL
```

```
all$Item_7 <- NULL
```

```
all$Item_8 <- NULL
```

```
all$Item_9 <- NULL
```

```
all$Item_10 <- NULL
```

```
all$Item_11 <- NULL
```

```
all$sum <- NULL
```

```
all$perc <- NULL
```

```
all$f1 <- NULL
```

```
all$f2 <- NULL
```

```
all$f3 <- NULL
```

```
all$f4 <- NULL
all$f5 <- NULL
all$mean <- NULL
view(all)

fit <- cfa(model1, data = all, estimator="MLR", mimic="Mplus")
summary(fit, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE)

####graphical Model:
semPaths(fit,whatLabels="std",edge.label.cex=1, style = "lisrel", residScale=8, layout
="tree3", theme = "colorblind", rotation= 2, what="std", nChartNodes = 0, curvePivot=
TRUE, sizeMan = 4, sizeLat = 10)

#####reliability analysis

#BUS 11 reliability = 0.867
alphaBUS <-data.frame(all$I1,all$I2, all$I3,all$I4, all$I5,all$I6,all$I7,all$I8,all$I9, all$I10,
all$I11)
cronbach.alpha(alphaBUS)

#F1= reliability = 0.906
alphaF1<-data.frame(all$I1,all$I2)
cronbach.alpha(alphaF1, standardized = TRUE, CI = TRUE)

#F2= reliability = 0.837
```



```
alphaF2 <-data.frame(all$I3,all$I4, all$I5)
```

```
cronbach.alpha(alphaF2, standardized = TRUE, CI = TRUE)
```

```
#F3= reliability = 0.856
```

```
alphaF3 <-data.frame(all$I6, all$I7,all$I8,all$I9)
```

```
cronbach.alpha(alphaF3, standardized = TRUE, CI = TRUE)
```