Artificial Intelligence Conversational Agents: Using Card Sorting To Revaluate The Validity Of The Chatbot Usability Scale

Bachelor Thesis

04-07-2022

Amber Ordelman

First supervisor: Jule Landwehr

Second supervisor: Simone Borsci

Abstract

This research aims to revaluate the validity of the Chatbot Usability Scale (BUS-11). This standardised tool was created to assess users' satisfaction with chatbots. The scale is revaluated in this research by testing its validity from a different angle, namely the face and construct validity. By doing so, the scale can be further validated.

As previous research has shown the importance of further investigating the effect of trust, this factor is considered in this study. It is tested whether trust might affect the mental models of participants. The results of testing the effect of trust on the mental models of participants might help future research to account for this factor. By doing so, outcomes might be more valuable.

Participants were asked to interact with two different chatbots. They had to fulfil two separate tasks, fill in the BUS-11 to assess their satisfaction, and fill in a trust questionnaire for each chatbot to assess their trust. To gain insight into the mental models of participants, a closed card sorting test was conducted. In order to check for the face and construct validity of the scale, heatmaps and item-level agreement matrixes were generated and then analysed. To qualitatively inspect the effect of trust on mental models, additional heatmaps and an item-level agreement matrix were created considering the trust levels of the participants. Moreover, to quantitively test the possible effect of trust on card sorting an Kruskal-Wallis test was performed.

The results of this research show that the mental models of participants appear to be in line with the expected item organization and the factorial structure of BUS-11. In addition, the previously found correlation between factor 2 and factor 3 of BUS-11 is also well reflected in the results of the card sorting. Moreover, trust appears to not have a significant effect on the mental models of the participants.

Thus, this research contributed to the ascertain the quality of the BUS-11 by testing its face and construct validity.

Keywords: Chatbots. User satisfaction, face validity, construct validity, BUS-11

Table of contents

1.Introduction

Technology is increasingly used by society and has been adopted into the daily lives of many humans worldwide. It has an impact on many aspects of daily life. To name a few, technology has impacted our access to information, has improved our communication and can relieve us from labour by taking over tasks at the workplace (Turner, 2022). Technology is always developing based on the society's demands and our way of living.

People make use of technology at school, at work, and during their spare time. In all domains, technology certainly has advantages. For example, technology in education created the opportunity for students to be surrounded by an engaging environment by incorporating different learning styles according to the students' needs (Walden University, 2022). At work, technology enhances the productivity of staff members and can create the opportunity to better help customers via faster and personalized customer service (Protected Trust, 2020). As for spare time, technology might enhance the quality of leisure. That is because it enables people to connect to distant areas easily, communicate faster and, it has brought new equipment that can be used for fun activities. Thus, the increasing use of technology has brought many advantages to society.

As society has become intertwined with technology, the human perspective is important to take into consideration. There are several factors that might affect the way humans interact with technology. One of these factors is trust which can be defined as a person's inclination to depend on another party due to its attributes (McKnight et al., 2011). Previous research has shown that trust can affect technology use: the more one trusts technology, the more one is likely to make use of the technology. Additionally, studies point out the importance of the effect of trust to be further analysed (McKnight et al., 2011). The relationship between trust and technology will become more important as the number of technological devices used in society increases.

One example of a domain in technology that has become increasingly important is Artificial Intelligence (AI). AI is a relatively new field of computer science that rises from the 50s. It is defined as the engineering and science of designing intelligent machines, in particular smart computer programs. AI is said to use computers to gain insight into human intelligence (Solutions, 2020). In essence, AI tries to mimic human higher functioning and can so be used to guide and assist humans. Our routines have changed immensely due to robotics and AI that are used in a wide variety of everyday services (Gabbay et al., 2009). AI can be used in several domains such as health care and customer services. In these domains, AI can have great advantages as it can relieve employees from extra labour. AI can be used to take over 'human tasks' such as administrative processes and answering questions of customers as a part of customer services. One form of AI that is often used in the domain of customer services are so called chatbots.

These chatbots, also referred to as conversational agents, are intelligent conversational applications that can mimic a human conversation by engaging in voice and/or text output and input (Borsci et al., 2021). One popular chatbot that is often used worldwide is Siri (Apple, 2007). When communicating with a chatbot, the application will go through three steps. Firstly, the chatbot will use, if available, conversation data to understand what kind of question you are asking. Secondly, the chatbot will analyse the right response to this question via a so-called training period. And lastly, the bot uses NLP (neuro-linguistic programming) and machine learning to learn context, and to improve its answers to similar questions asked in the future (Porter, 2022). There are three types of chatbots that are mostly used nowadays: live-chat, rules-based chatbots and AI chatbots (Porter, 2022). These chatbots can serve different purposes in the previously mentioned AI domains but might also serve for entertainment, website help and education (Valtolina, Barricelli, Gaetano & Diliberto, 2018). Thus, chatbots are not only widely deployable and can positively impact our lives, but according to research are also predicted to make up 85% of customer interaction in the future (Borsci et al., 2021).

Despite these forecasts, there appeared to be a lack of knowledge regarding the end-user satisfaction with chatbots. The satisfaction of users can be defined as the comfort and attitude one has towards the technology that is used. This satisfaction can be measured via attitude rating scales (Frøkjær, Hertzum & Hornbæk, 2000). It is of great importance to assess the user's satisfaction for several reasons. Firstly, the concept of usability and satisfaction are related. Usability can be described as the degree to which a specified type of user is able to reach a specified goal with satisfaction, effectiveness, and efficiency by making use of a product (International Organization for Standardization [ISO], 2018). Satisfaction is an element of usability and of user experience. Within the context of usability, satisfaction may only result from using the product (ISO, 2018). Logically, the user of the chatbot should experience a high level of usability as the chatbot should help the user by for example answering a question and hence reaching a specific goal. As satisfaction is one way to measure usability, gaining knowledge on this aspect of chatbots can be useful. Research has shown that a higher level of usability has a positive influence on the users' satisfaction (Gocardless Team, 2021). Secondly, satisfaction is related to a users' greater website loyalty (Gocardless Team, 2021). Phrased differently, a high satisfaction level will more likely result in the user returning to the website. Lastly, when a company has invested in a chatbot, its goal will likely be to serve their customer with satisfaction as the end goal. To reach that goal, one should measure satisfaction and improve certain hiccups that appear from the measurement.

Taking the above-mentioned information into consideration, a lack of knowledge regarding the end-user satisfaction might be viewed as problematic. Hence, the 'chatBot Usability Scale' (BUS) was developed to enable users of chatbots to express their satisfaction level. In addition, the scale creates the possibility for standardized measurement, enabling evaluators and designers to compare their outcomes (Borsci et al., 2021). As described in the paper of Borsci and colleagues, the new scale was designed by making use of a systematic literature review, three other studies with an overall sample of 141 participants in the survey, focus group sessions and testing of chatbots. Additionally, the scale has been revised after it was tested. The first version of the scale consisted of 15 items which emerged from the exploratory analysis. After doing confirmatory factor analysis, 4 items were deleted from the scale resulting in a final version of 11 items with five factors: BUS-11. This revised scale is the focus of this research. All factors and matching items of BUS-11 are displayed below in Table 1.

Table 1

The original factorial structure of the BUS-11. All five factors are represented in the left column and all 11 items in the right column.

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	6	I find that the chatbot understands what I want and helps me achieve my goal.
	7	The chatbot gives me the appropriate amount of information.

		8	The chatbot only gives me the information I need.
		9	I feel like the chatbot’s responses were accurate.
4	- Perceived privacy and security	10	I believe the chatbot informs me of any possible privacy issues.
5	- Time response	11	My waiting time for a response from the chatbot was short.

As for the psychometric properties of the scale, the internal consistency of BUS-11 and its factors was tested by using Cronbach’s alpha. The outcome showed a high internal consistency (α = .89), meaning that the scale was proven to be reliable (Huijsmans, 2022). Moreover, research has shown that BUS-11 measures the satisfaction of users with chatbots, showing validity. In addition, a strong correlation was found between factor 2 ‘perceived quality of chatbot functions’ and factor 3 ‘perceived quality of conversation and information provided’ (Borsci et al., 2021). This correlation could be explained as both these factors appear to measure a form of perceived quality (Huijsmans, 2022). Nevertheless, a strong correlation between factors might affect the validity of a scale.

For a scale to be optimally effective, it should be both reliable and valid. As described above, BUS-11 was proven to be reliable by research before. Hence, the validity of the scale is the main point of focus in this research. This study aims to provide insight regarding its validity by using the method of card sorting. Card sorting can be used to gain insight into how users structure and classify content. This established method was chosen as it enables measuring the validity of the scale. Card sorting helps gaining insight into the mental models of participants (Nawaz, 2012). By doing so, it unravels how information should be organized according to the participants and it can be tested whether their mental models are similar to the original factorial structure of BUS-11. Card sorting can be done both open and closed. In this research, closed card sorting is used as this creates the possibility of using predefined categories: the constructs of the scale (see Table 1). Participants are asked to match each item to a category, creating the possibility to gain insight into how participants match each item to a construct (Assistant Secretary of Public Affairs, n.d.). The scale is designed in a particular manner so that each item belongs to a certain underlying factor (see Table 1). Hence, the way participants do the card sorting provides insight into whether participants feel like the items measure the constructs as originally intended. This can provide information about the validity of the scale.

The term validity refers to whether one is accurately measuring what one wants to measure. In this research, two types of validity will be the point of focus: face validity and construct validity. Firstly, face validity refers to a subjective assessment of whether an item is a good measure or not (Fitzner, 2007). Thus, in this case, whether a particular item in the BUS-11 appears to belong to a specific factor in line with the original construct according to the participants. After obtaining data from the card sorting test, the face validity was tested by assessing the number of participants that has grouped each item to their original construct (Beerlage-de Jong, Kip & Kelders, 2020). Secondly, construct validity addresses if the scale accurately measures the construct that it is expected to measure (Middleton, 2022). In our case, if there is a match or not between the original construct and the one that emerged from the card sorting. To test the construct validity, the card sorting data was analysed to check how participants have grouped items and how they have matched the items to the constructs in line with their mental models (Beerlage-de Jong, Kip & Kelders, 2020).

Aside from testing the validity of the scale, this research will test the possible effect of trust on the mental models of participants and hence whether they are able to make sense of the original factorial structure of BUS-11. As mentioned earlier, previous studies have shown the importance of further investigating the effect of trust. Hence, this factor is included in the research.

To conclude, this work aims at further validating the BUS-11 scale by using card sorting. By doing so, the face and construct validity of the BUS-11 can be investigated. Hereby, the aim is to optimize the satisfaction scale. Moreover, to explore the possible effects of trust towards technology on the mental models of participants, their level of trust is assessed and considered. By doing so, knowledge can be gained regarding the possible effect of trust on mental models of participants.

In order to make statements about the face and construct validity and to assess the possible effect of trust, two research questions will be answered:

RQ1: Are the mental models of participants matching the relation between the items and factors of BUS-11 as originally intended?

RQ2: Is trust affecting the outcomes of the card sorting test?

2.Methods

2.1 Participants

Before starting, the participants have signed the informed consent form (Appendix A). The research necessitated 55 participants in total and participants who did not complete the survey were excluded. As a result, 23 were included in the data analysis. The age ranged from 17 to 53. Regarding the current gender identity, 9 identified as man and 13 as woman. Additionally, 1 participant declined to answer. The level of English proficiency is B1-intermediate for three participants, B2-upper for six, C1-advanced for 10 participants, and C2-proficient for four participants. The sampling method for this study was convenient sampling. Additionally, the research has been approved by the BMS committee.

2.2 Materials

The experience management software program Qualtrics was used to create the online test that participants had to fill in. This program was chosen as online testing might provide a larger and more diverse sample. The template from the University of Twente was used to create the informed consent form (Appendix A).

2.2.1 Demographical questions

The participants were asked to answer four demographical questions to get a general view of their characteristics. The questions are as follows; ‘How old are you?’, ‘What is your current gender identity?’, ‘What is your nationality?’, and ‘What is your English proficiency?’.

2.2.2 Chatbots

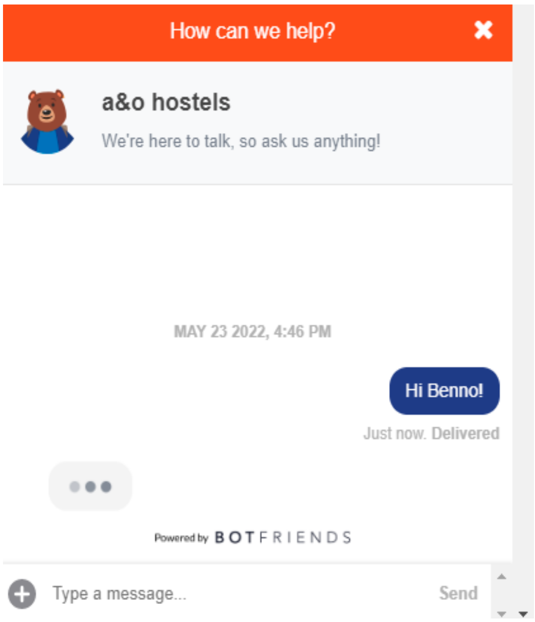
Two chatbots are used to make participants more familiar with this type of technology. Additionally, participants are asked to engage in a certain task for each chatbot. By giving participants these tasks, they must interact with the chatbots. At the end, participants are asked to answer a question they are only able to fill in after they have interacted with the bot. By doing so, it can be checked whether the participants engaged in interaction with the chatbots. It might be the case that even though participants have tried, they were not able to find the answer. Hence, it is asked whether they were able to complete the task.

2.2.3 Chatbot 1 A&O Hostels

The first chatbot presented to the participants is the A&O Hostel chatbot as can be seen in Figure 1. The participants must answer how much a parking ticket costs in Berlin (Friedrichshain). By interacting with the chatbot, participants can find the answer.

Figure 1

The A&O hostels chatbot participants are asked to interact with. This is the first chatbot represented to the participants.



How can we help?

a&o hostels

We're here to talk, so ask us anything!

MAY 23 2022, 4:46 PM

Hi Benno!

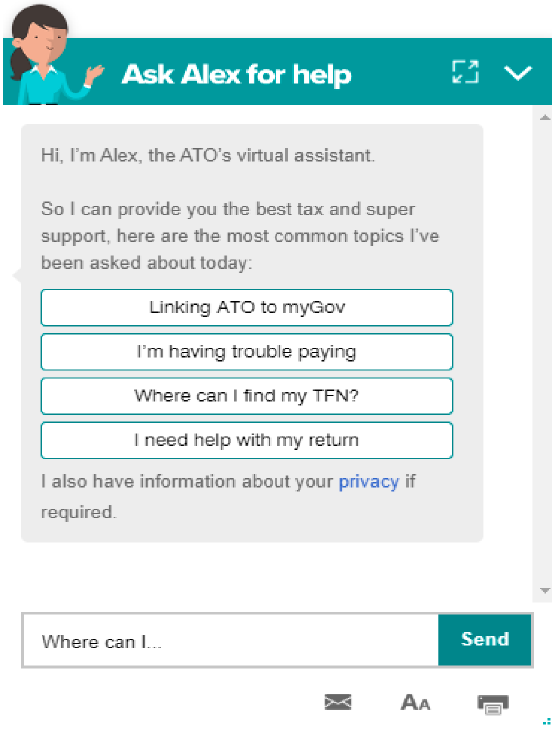
Just now. Delivered

2.2.4 Chatbot 2 ATO

The second chatbot that is presented to the participants is that of ATO (see Figure 2). By interacting with this chatbot, participants must find out when the deadline is to submit the tax return when doing it yourself.

Figure 2

The ATO chatbot participants are asked to interact with. This is the second chatbot represented to the participants.



Hi, I'm Alex, the ATO's virtual assistant.

So I can provide you the best tax and super support, here are the most common topics I've been asked about today:

I also have information about your privacy if required.

Where can I...

2.2.5 Satisfaction questionnaire

The BUS-11 designed by Borsci and colleagues that can be used to assess the user's satisfaction with chatbots was implemented in the test (Appendix B). This scale exists of 11 items all addressing aspects of usability. These items are presented to the participants as statements, and they have to rate these statements by making use of a 5-point Likert scale.

This means that participants can rate the item from strongly disagree to strongly agree. The 11 items presented to the participants all measure an underlying construct. An overview of the items and their constructs can be found in Appendix C. BUS-11 itself was implemented in the test as participants must engage in a card sorting test for the scale at the end. Hence, when they use it themselves, they might be better able to evaluate the scale.

2.2.6 Trust questionnaire

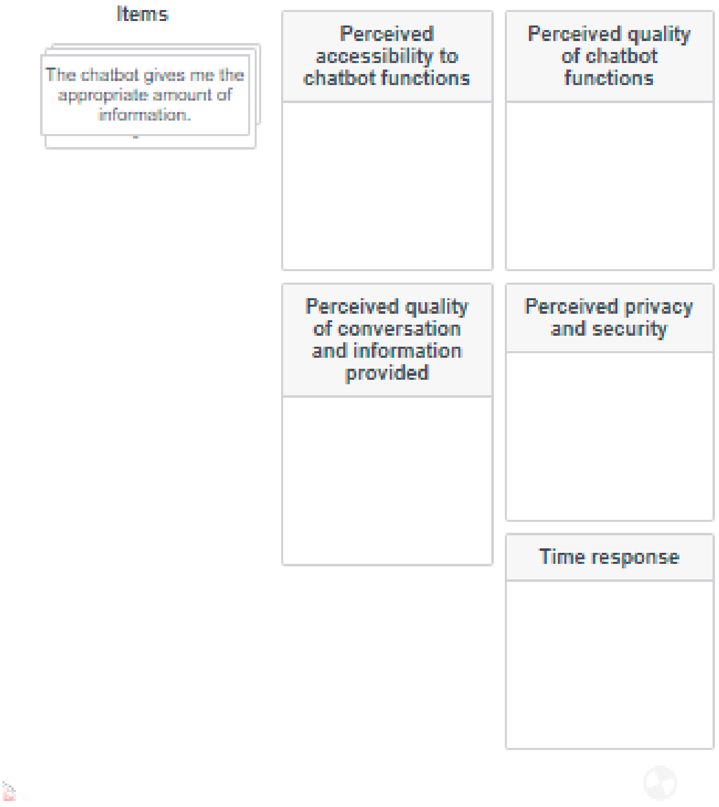
To assess the level of trust of the participants, the trust questionnaire created by McKnight was implemented (Appendix D). This particular questionnaire was used as it is an evaluated and standardized scale. The questionnaire contains 20 items in total and all items are divided over four different categories. The categories of the questionnaire are trust, expertise, human-likeness, and risk. The items are presented as statements and participants are asked to assess their agreement on a 7-point Likert scale ranging from strongly disagree to strongly agree. For example, they are asked to indicate whether they experience the chatbot as trustworthy. After the participants have interacted with a chatbot and have filled in the satisfaction scale, the trust scale is presented. In the title of the trust scale, it is made clear that participants should fill it in for the chatbot they have just used (either A&O or ATO).

2.2.7 Card sorting

A closed card sorting test was used in the Qualtrics test. This particular test is used to revaluate the chatbot satisfaction scale by gaining insight into the mental models of the participants. Figure 3 shows how the card sorting is presented to the participants. The 11 items of the satisfaction scale are displayed on the left and the five constructs on right: ‘Perceived accessibility to chatbot functions’, ‘Perceived quality of chatbot functions’, ‘Perceived quality of conversation and information provided’, ‘Perceived privacy and security', and 'Time response'. Participants are instructed to drag the item to the construct they think it belongs to.

Figure 3

The closed card sorting test presented to the participants. The 11 items of the BUS-11 can be found on the left and the underlying factors are displayed on the right.



Items
Perceived accessibility to chatbot functions
Perceived quality of chatbot functions
Perceived quality of conversation and information provided
Perceived privacy and security
Time response

The chatbot gives me the appropriate amount of information.

2.3 Procedure

Respondents were first informed about the study and their participation through the informed consent form. After the participant had carefully read through the document, the informed consent was signed online. After gaining consent, the participants had to fill in four demographical questions. Next, participants were given a short definition and an example of a chatbot. This was done to help all participants to understand what a chatbot entails. After this, questions regarding the first chatbot (A&O Hostels) were shown. The participants were asked to go to the website as described on their online screen. They were asked to fulfil a certain task: finding out what a parking space costs in Berlin (Friedrichshain). They were asked to indicate the price afterwards and were asked whether they were able to complete the task. Next off, they had to fill in the chatbot usability scale for the A&O Hostels chatbot. After completing the usability scale, they were asked to answer the trust questionnaire for the A&O Hostels chatbot. The same process was repeated for the second ATO chatbot; task chatbot, satisfaction scale and then trust questionnaire. The task for this chatbot was to find out when the deadline is to submit/lodge their tax return when doing it yourself. After this, the participant was asked to engage in a closed card sorting test. In this test, the participants were asked to match all items on the scale to a certain factor according to what they feel like is most logical.

2.4 Data analysis

After gathering the data of 55 participants, the data was analysed. Data from participants who did not complete the test were removed from the data set, resulting in a total of 23 usable responses. Now that the final data set is determined, the data analysis can start.

To address RQ1: '*Are the mental models of participants matching the relation between the items and factors of BUS-11 as originally intended?*', the construct validity and face validity are being tested. Firstly, the analysis done for construct validity will be discussed. As the aim is to detect clusters of items made by the participants, a cluster analysis is done. At first, a general similarity matrix was created with the data of all 23 participants by using the Jaccard coefficient. This coefficient gives insight into the similarity between two objects. In the case of this research, it shows the similarity between two items in the card sorting (Schmettow & Sommer, 2016). This coefficient is calculated by dividing two calculations and is done for every participant individually. The first calculation entails the number of groups both items were grouped in (intersection set). For example, if a participant groups item 1 and item 2 together into the same group, the number of the intersection set goes up with 1. The second calculation addresses the number of groups either of the two items were grouped in (union set). For example, if a participant groups item 1 in a group without item 2, the number of the union set still goes up with 1. Next off, the number of the intersection set should be divided by the number of the union set (Schmettow & Sommer, 2016). After calculating every coefficient, the results are combined in a similarity matrix. Next, a heatmap was created in R to visualize the matrix. This was done by a vector analysis. The data was first transformed in numerical format, names were given and colors of the heatmap were defined (red for high numbers). The heatmap is then analysed in order to detect clusters of items. Hereby, insight can be gained regarding the construct validity of the scale.

For the face validity, an item to factor matrix was created by item-level agreement. Item-level agreement is determined by assessing how many participants have grouped each item in the original intended construct. This is calculated by checking how many times an item has been matched to the intended factor. Next off, that number will be divided by the ideal number of times an item has been matched to its factor (in this case 23). This is done for every factor and will give a certain percentage that shows how often participants have matched an item to a certain factor (Beerlage-de Jong, Kip & Kelders, 2020). Based on these percentages, it is determined whether there is a complete match, a partial match, or no match with the original construct. By analysing this matrix, statements can be made regarding the face validity of the scale.

In order to answer RQ2: '*Is trust affecting the outcomes of the card sorting test?*', the effect of trust is taken into account. To test this, data obtained from the trust questionnaires implemented in the Qualtrics test are used. This includes a total of 22 participants as one participant did not complete both trust questionnaires and is therefore excluded. First, to inspect the effect of trust, the participants were divided in three separate groups: a ‘average group’, a ‘below average group’, and a ‘above average group’. These groups are the results of scoring all answers of the 22 participants on the trust-scale with a 7-point Likert scale. Those scoring within 1 SD (σ=7.8) from the mean (μ=83.95) where included in the ‘average group’ (N=12). Those scoring more than 1 SD below the mean were placed in the ‘below average group’ (N=4) and those scoring more than 1 SD above the mean were placed in the ‘above average group’ (N=6).

Based on these groups, three separate similarity matrixes were created to qualitatively test the effect of trust. After these similarity matrixes were created, heatmaps were made to visualize the outcomes. The three different heatmaps were used to provide qualitative insights about the possible effect of trust on the construct validity of BUS-11. A summary of the match of the construct and the card sorting per trust group was created to visually inspect the relationship between trust and mental models.

Finally, a Kruskal-Wallis test was performed to quantitively test the effect of trust on card sorting. This test was conducted as the data appeared to not be normally distributed.

3. Results

3.1 Construct and face validity of BUS-11

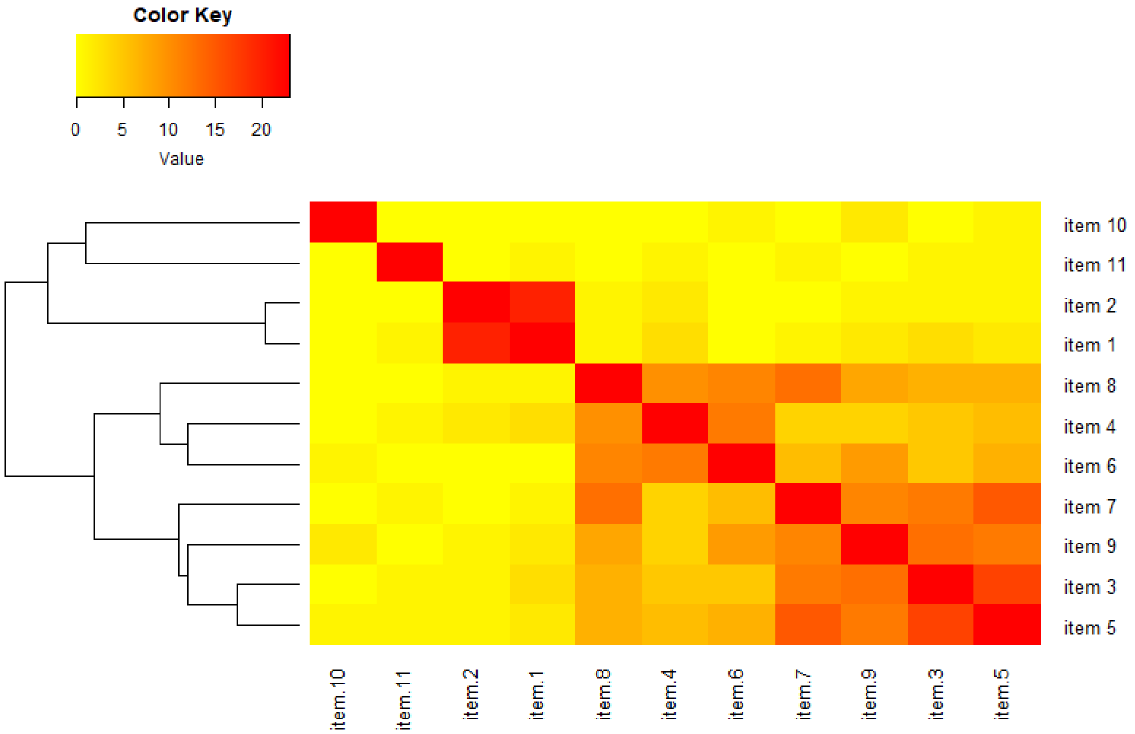
In order to explore the construct validity of BUS-11, a general similarity matrix was created as described in the data analysis. Figure 4 shows the general similarity matrix displayed as a heatmap. When observing Figure 4, at the bottom right of the diagonal more scattered colors appear. The more scattered distribution of colors shows more variation in the clusters of items made by the participants. In this scattered area, it can be seen that participants have frequently clustered the following items together:

a. Item 3 and item 9
b. Item 4 and item 6
c. Item 5 and item 7

In the original scale, these clustered items measure different constructs. Nevertheless, the associations made are represented in a dark orange color showing that these items could potentially correlate.

Figure 4

Heatmap for the general similarity matrix with the 11 items of the BUS-11 represented on the axis. The color key shown in the top left is ranging from 0 (yellow) to 23 (dark red) which represents the intensity of the relationship: 0 indicates that a certain match has not been made by the participants, whereas a value of 23 shows that a relationship was identified by all the participants.



This scattered distribution found in the bottom right of Figure 4, is in contrast with the top left of the diagonal. Here, the dark red shows the cluster of item 1 and item 2 made by the participants. The dark red is solely surrounded by yellow, meaning that the majority of participants have made similar clusters. The cluster of item 1 and item 2 is in line with expectations as these items intend to measure a similar construct in the original study of Borsci and colleagues: ‘perceived accessibility to chatbot functions’ (Appendix C). Additionally, participants have clustered the following items together:

- Item 3 and item 5
- Item 6 and item 8
- Item 7 and item 8

These clusters of items are in line with expectations as these items measure the same constructs in the original scale.

In order to gain insight into the face validity of the scale, a general item-level agreement matrix is made as shown in Table 2. When analysing this table, it can be observed that five of the 11 items are showing a complete match with the original factorial structure. A complete match was found for item 1, item 2, item 7, item 10, and item 11. This shows that participants can appropriately associate these items as belonging to the factor as intended in the original construct.

Table 2

General item-level agreement matrix displaying the matches made between the 11 items and the five factors of the BUS-11 by the participants (face validity). The original factors are displayed in the column on the left. In the last column on the right, it is indicated whether there is a complete match, a partial match, or no match with the original factorial structure (construct validity). A threshold of 75% was used to determine a complete match.

Original factor	**Item**	ACCES[a](%)	QUAL[b](%)	CONV[c](%)	PRIV[d](%)	TIME[e](%)	MATCH[d]
ACCES	1	91	9	0	0	0	YES
ACCES	2	91	0	0	0	9	YES
QUAL	3	4	35	61	0	0	PARTIAL
QUAL	4	9	61	26	4	0	PARITAL
QUAL	5	4	22	74	0	0	PARTIAL
CONV	6	4	70	30	0	0	PARTIAL
CONV	7	4	9	87	0	4	YES
CONV	8	9	39	57	0	0	PARTIAL
CONV	9	9	35	61	0	0	PARTIAL
PRIV	10	0	4	4	91	0	YES

TIME	11	0	4	0	0	96	YES

ACCES[a] = Perceived accessibility to chatbot functions

QUAL[b] = Perceived quality of chatbot functions

CONV[c] = Perceived quality of conversation and information provided

PRIV[d] = Perceived privacy and security

TIME[e] = Time response

MATCH[d] = Match with original factorial structure

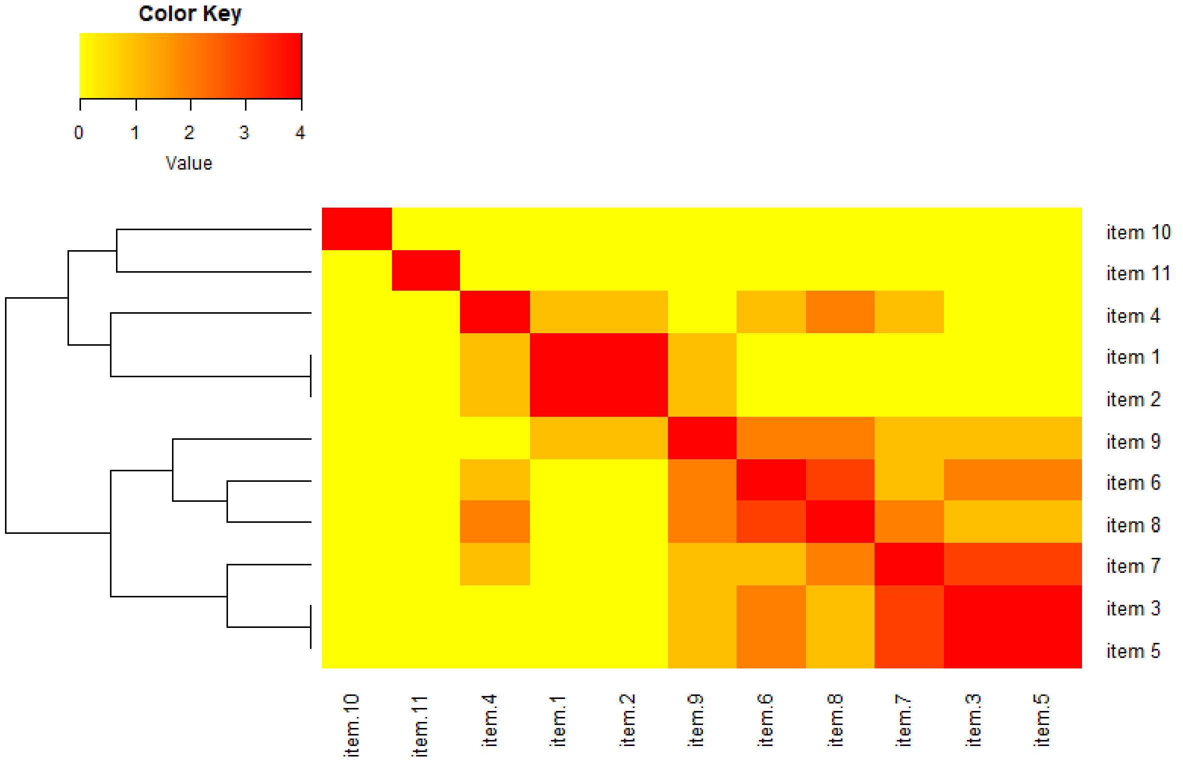
When further observing Table 2, it can be seen that two items in particular were frequently matched to a different factor than originally intended. This accounts for item 6 that was 70% of the time matched to factor 2 (QUAL[b] in Table 2) and item 5 that 74% of the time matched to factor 3 (CONV[c] in Table 2). These percentages are just below the threshold and hence are labelled as a partial match. Nevertheless, these relatively high percentages show it was more challenging for participants to appropriately associate these items as belonging to their factors as proposed in the original study. This might affect the face validity.

3.2 The effect of trust on the outcome of the card sorting

To inspect for the effect of trust on the construct validity, separate heatmaps were created as explained in the data analysis. In Figure 5, the heatmap for the below trust group can be found.

Figure 5

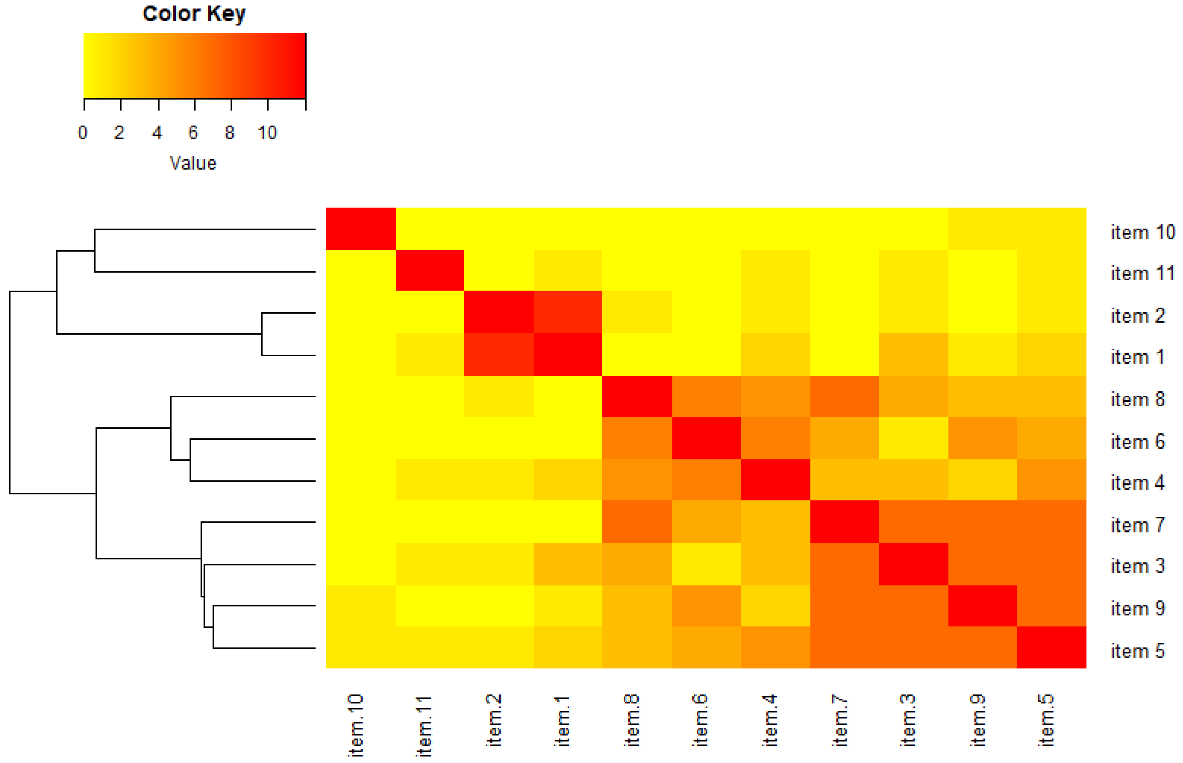
Heatmap for the separate 'below average' trust group with the 11 items of the BUS-11 represented on the axis. The color key shown in the top left is ranging from 0 (yellow) to 4 (dark red) as four participants are included in this group. This key represents the intensity of the relationship from 0 indicating that a certain match has not been made by the participants *whereas a value of 4 shows that a relationship was identified by all the participants.*



The possible difference in between the separate groups is checked when analysing the heatmaps in Figure 5, Figure 6, and Figure 7. When observing these heatmaps, there is a broader distribution of colors in the ‘above average’ group and ‘average’ group in contrast with the ‘below average’ group’. That means that the participants in the ‘below average’ group more frequently clustered the same items together. This can be observed in the heatmap in Figure 5 by more darker red and less variation in orange and yellow tones as opposed to the heatmap of the ‘average’ group in Figure 6.

Figure 6

Heatmap for the separate ‘average’ trust group with the 11 items of the BUS-11 represented on the axis. The color key shown in the top left is ranging from 0 (yellow) to 12 (dark red) as 12 participants are included in this group. This key represents the intensity of the relationship from 0 indicating that a certain match has not been made by the participants whereas a value *of 12 shows that a relationship was identified by all the participants.*

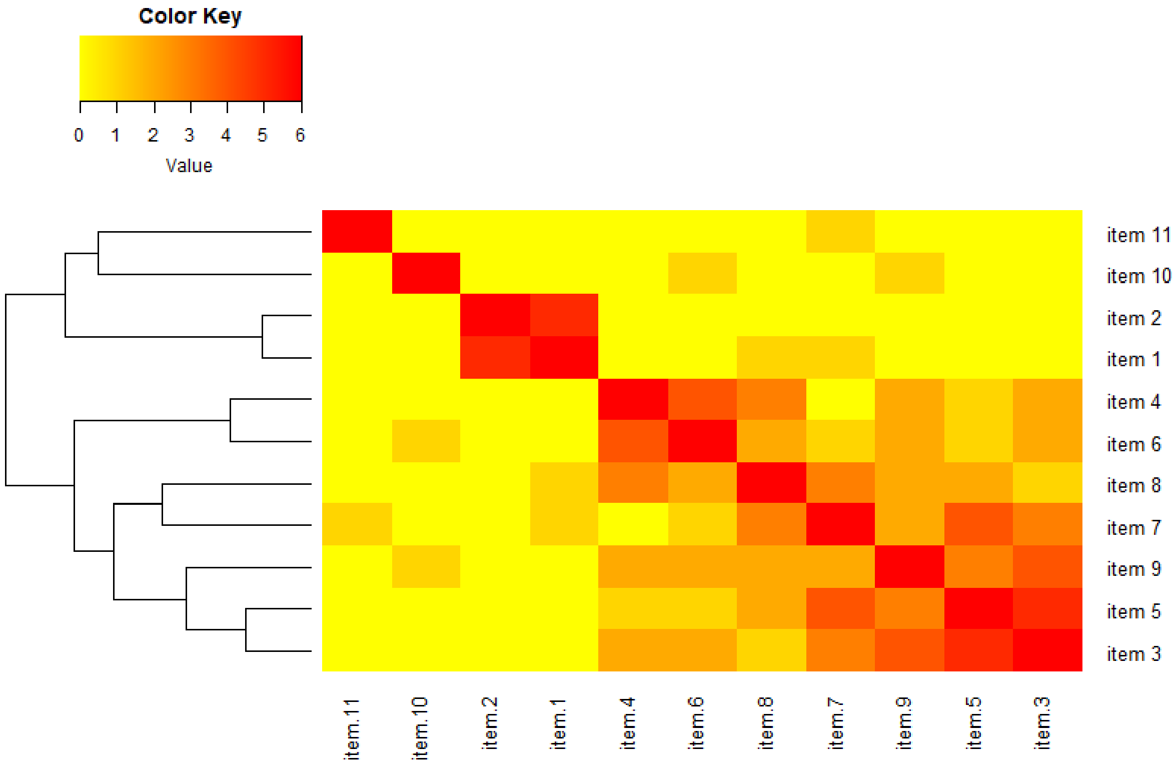


Aside from the difference in distribution of clusters found, there can also be differences found in the clusters of items made by the participants of the different groups:

- Clusters of items in 'average group': item 3 and item 7, item 5 and item 7, item 5 and item 9, item 3 and item 9
- Clusters of items in 'below average' group: item 4 and item 8, item 3, and item 7
- Clusters of items in 'above average' group: item 4 and item 6, item 4 and item 8, item 9 and item 3, and item 5 and item 7

Figure 7

Heatmap for the separate 'above average' trust group with the 11 items of the BUS-11 represented on the axis. The color key shown in the top left is ranging from 0 (yellow) to 6 (dark red) as 6 participants are included in this group. This key represents the intensity of the *relationship from 0 indicating that a certain match has not been made by the participants whereas a value of 6 shows that a relationship was identified by all the participants.*



Comparing the clusters as presented above Figure 7, it can be observed that the groups agree regarding some clusters (item 3 and item 9, and item 5 and item 7). However, they differ when looking at others. As the three groups differ in their level of trust, it might mean that this factor affects how they approach matching certain items together. This in turn might affect the construct validity of the scale.

In addition, an item-level agreement matrix with different trust groups was created. As can be seen in Table 3, a relatively high item-level agreement can be found at certain item-factor combinations. The table shows that all groups have matched the following items to their original factor: item 1, item 2, item 10, and item 11. Visually it seems that the different levels of trust do not significantly affect the outcomes compared to the overall group of participants. The only difference that can be found is that the 'above average trust' group have made a partial match on item 7 and factor 3.

Table 3

Item-level agreement matrix displaying the 11 items and five factors of the BUS-11. In the table itself, it is indicated for each group whether there is a complete match, a partial match, or no match with the original factorial structure (construct validity). In the right column, the matches can be found made by all participants without accounting for their level of trust. A threshold of 75% was used to determine a complete match.

Original factor	**Item**	**Match with original Factorial structure**			
		Above	Average	Below	Item level agreement
ACCES	1	Match	Match	Match	Match
ACCES	2	Match	Match	Match	Match
QUAL	3	Partial	Partial	Partial	Partial
QUAL	4	Partial	Partial	Partial	Partial
QUAL	5	Partial	Partial	Partial	Partial
CONV	6	Partial	Partial	Partial	Partial
CONV	7	Partial	Match	Match	Match
CONV	8	Partial	Partial	Partial	Partial
CONV	9	Partial	Partial	Partial	Partial
PRIV	10	Match	Match	Match	Match
TIME	11	Match	Match	Match	Match

ACCES = Perceived accessibility to chatbot functions

QUAL = Perceived quality of chatbot functions

CONV = Perceived quality of conversation and information provided

PRIV = Perceived privacy and security

TIME = Time response

To further inspect the possible differences between the trust groups, regression analysis is done. The outcomes of the Kruskal-Wallis test show a non-significant effect, H (2) = 0.16, p = 0.92. This means that there appears to be no significant effect of the level of trust on the ability of the participants to appropriately recognize the relation between the items and factors of BUS-11 as originally intended.

4. Discussion

This research contributed to optimizing the BUS-11 by revaluating the validity of the scale. This is done by testing the face and construct validity of BUS-11. It is of great importance to evaluate the scale as a valid and reliable scale is essential for gaining the needed knowledge regarding the satisfaction of chatbot users. Validity is revaluated as previous research has found factors 2 and 3 to be highly correlated and the scale has been proven to be reliable (Borsci et al., 2021); (Huijsmans, 2022). In addition, it was checked whether trust influences the match of the mental model of participants with the original intended structure of BUS-11. The effect of trust was tested as previous research has pointed toward the importance of trust to be further analysed (McKnight et al., 2011).

To answer our first research question, '*Are the mental models of participants matching with the relation between the items and factors of BUS-11 as originally intended?*', it appears that the construct validity of the scale can be confirmed. That is because participants generally clustered items together that measure the same original construct. Moreover, participants made three other frequent clusters that do not measure the same original construct. Nevertheless, this can be explained as all items included in these clusters either belong to factor 2 or factor 3 and these factors were found to be highly correlated. Additionally, participants were able to correctly match most items to their original construct. The partial matches found can be explained as these items belong either to factor 2 or factor 3. As mentioned before, these factors highly correlate, explaining the partial matches made by participants. In addition, it appears that face validity can be confirmed as well. That is because the majority of the participants was able to group the individual items in their original construct. The items that were frequently grouped to a different construct all belong to either factor 2 or factor 3. Hence, these combinations can be explained. Thus, it appears that the mental models of participants are matching the relation between the items and factors of BUS-11 as originally intended.

In addition, the results of this study confirm the strong association between factor 2 and factor 3 as shown before by Borsci and colleagues (2021) and Huijsmans (2022). It might be that these factors do not sufficiently differ from each other, and future research should further investigate whether they can be merged together. That is because a high correlation as found before between factor 2 and factor 3 might be viewed as problematic (Rönkkö & Cho, 2020).

The second research question was: '*Is trust affecting the outcomes of the card sorting test?* Regarding this question, it can be stated that there might be an effect solely based on the qualitative analysis. That is because the level of trust appears to influence the clusters of items that participants made. Additionally, it appears that those having a 'below average' trust level agree more frequently on the clusters that were made. This might assume that the mental models of these participants are more similar in comparison with the 'above average' and 'average' groups. However, further analysis has proven this effect is non-significant. Hence, it can be concluded that trust is not affecting the results of the card sorting outcome despite the match of the original constructs in BUS-11. In addition, it appears that the participants of all three groups were generally making the same matches. This means that the level of trust does not affect the level of agreement with the original constructs in BUS-11. This conclusion is confirmed by the non-significant difference found in further quantative analysis.

4.1 Limitations and recommendations for future research

Despite the fact that this research has resulted in some valuable knowledge, there are four limitations that should be considered. Firstly, the sample size of the separate trust groups can be considered a limitation. As the groups consist of a rather small number of participants, the generalizability of the results with regard to the effect of trust might be limited. Increasing the sample size might affect the outcome of the quantitative analysis. This is because a larger sample size increases the chance on a higher significance level of the outcomes (Kalla, 2009).

Additionally, the participants had a limited age range as the research mainly made use of adolescents. It might be that outcomes would have been different if more elderly participants were included as previous research has shown a difference in trust levels between different age groups. It was found that elderly, generally speaking, have less experience with technological devices and this might affect their level of trust (SafeHome.org Team, 2015). Therefore, regarding both the limitation of trust and sample size, future research should make use of a larger and more diverse sample size regarding age.

Moreover, participants had different levels of English proficiency as filled in during the test. This might have affected the answers they filled in and hence the results of the research.

In addition, the card sorting method can have a disadvantage. This method does not enable the researcher to get an understanding of the reasoning behind the decisions that a participant makes during the card sorting process (Beerlage-de Jong, Kip & Kelders, 2020). When taking use of interviews for example, one could gain insight into these decisions. This might lead to more valuable information.

5. Conclusion

This study has contributed to the further validation of the BUS-11 by testing its face and construct validity. The results have confirmed both the face and construct validity of the scale as those making use of the BUS-11 were able to recognize the relation between the items and factors as originally intended. Moreover, the strong correlation between factor 2 and factor 3 can be confirmed. More research is needed to investigate whether it would be better to merge these factors together. Additionally, a non-significant effect of trust on the mental models of participants was found. To conclude, this revaluation of the BUS-11 has confirmed this scale to be valid. So, the scale measures users satisfaction with chatbots.

6 References

Assistant Secretary for Public Affairs. (n.d.). Card Sorting | Usability.gov. Usability.gov. Retrieved 15 March 2022, from https://www.usability.gov/how-to-and tools/methods/card-sorting.html

Apple. (2007). *Siri*. Retrieved 12 June 2022, from https://www.apple.com/siri/

Beerlage-de Jong, N., Kip, H., & Kelders, S. M. (2020). Evaluation of the Perceived Persuasiveness Questionnaire: User-Centered Card-Sort Study. *Journal of Medical Internet Research*, *22*(10), e20404. https://doi.org/10.2196/20404

Borsci, S., Malizia, A., Schmettow, M., Van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2021). 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

Fitzner, K. (2007). Reliability and Validity A Quick Review. *The Diabetes Educator*, *33*(5), 775–780. https://doi.org/10.1177/0145721707308172

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00*. https://doi.org/10.1145/332040.332455

Gabbay, D. M., Thagard, P., Woods, J., & Meijers, A. W. M. (2009). *Philosophy of Technology and Engineering Sciences*. Elsevier Gezondheidszorg.

Gocardless Team. (2021, 12 april). *Why is customer satisfaction so important?* GoCardless. Retrieved 12 June 2022, from https://gocardless.com/guides/posts/customer satisfaction/

Huijsmans, M., (2022) *The chatbot usability scale : an evaluation of the Dutch version of the BUS-11.*

ISO (2018), ISO 9241-11:2018. Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts

Kalla, S. (Jun 18, 2009). Statistical Significance And Sample Size. Retrieved Jun 13, 2022 from Explorable.com: https://explorable.com/statistical-significance-sample-size

The Importance of Modern Technology in the Workplace - Protected Trust – Everything Microsoft 365 and Surface for your business. (2020, 1 mei). Protected Trust Everything Microsoft 365 and Surface for Your Business. Retrieved 12 June 2022, from https://www.protectedtrust.com/technology-in

theworkplace/#:%7E:text=It%20can%20provide%20higher%2Dquality,workplace%20creates%20a%20competitive%20advantage.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. *ACM Transactions on Management Information Systems*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Middleton, F. (2022, 24 februari). *The four types of validity*. Scribbr. Retrieved 18 March 2022, from https://www.scribbr.com/methodology/types-of-validity/

Nawaz, A. (2012). A Comparison of Card-sorting Analysis Methods. In APCHI '12. Proceedings of the 10th Asia Pacific Conference on Computer-Human Interaction (Vol. 2, pp. 583- 592). Association for Computing Machinery.

Porter, E. (2022, 22 maart). *Chatbot*. Drift. Retrieved 12 June 2022, from https://www.drift.com/learn/chatbot/

Rönkkö, M., & Cho, E. (2020). An Updated Guideline for Assessing Discriminant Validity. *Organizational Research Methods*, *25*(1), 6–14. https://doi.org/10.1177/1094428120968614

SafeHome.org Team. (2015, 5 februari). *Privacy And Technology*. SafeHome.Org. Geraadpleegd op 13 juni 2022, van https://www.safehome.org/resources/privacy-and-technology/

Schmettow, M., & Sommer, J. (2016). Linking card sorting to browsing performance – are congruent municipal websites more efficient to use? *Behaviour &amp; Information Technology*, *35*(6), 452–470. https://doi.org/10.1080/0144929x.2016.1157207

Solutions, A., Brown, G., Brown, G., & Brown, G. (2020). *Homage to John McCarthy, the Father of Artificial Intelligence (AI) Page 1 of 0*. Conversational AI Platform for Enterprise - Teneo | Artificial Solutions. Retrieved 12 June 2022, from https://www.artificial-solutions.com/blog/homage-to-john-mccarthy-the-father-of artificialintelligence#:%7E:text=It%20was%20in%20the%20mid,engineering%20of 20making%20intelligent%20machines%E2%80%9D.

Turner, J. (2022, 6 mei). *The 7 Main Ways Technology Impacts Your Daily Life*. Tech.Co. Retrieved 12 June 2022, from https://tech.co/vpn/main-ways-technology impacts-daily-life

The Importance of Modern Technology in the Workplace - Protected Trust - Everything Microsoft 365 and Surface for your business. (2020, 1 May). Protected Trust -- Everything Microsoft 365 and Surface for Your Business. Retrieved 12 June 2022, from https://www.protectedtrust.com/technology-in-the-workplace/#:%7E:text=It%20can%20provide%20higher%2Dquality,workplace%20creates%20a%20competitive%20advantage.

Valtolina, S., Barricelli, B.R., Gaetano, S.D., & Diliberto, P. (2018). Chatbots and Conversational Interfaces: Three Domains of Use. *CoPDA@AVI*.

Walden University. (2022, 31 mei). *Top 5 Benefits of Technology in the Classroom*. Retrieved 12 June 2022, from https://www.waldenu.edu/programs/education/resource/top-five-benefits-of technology-in-the-classroom

8. Appendices

Appendix A

The Participation information sheet and informed consent form

UNIVERSITY OF TWENTE.

University of Twente
Bachelor Programme in Psychology

Participation Information Sheet

Artificial Intelligence Conversational Agents: Using Card Sorting To Evaluate The Chatbot Usability Scale'

What is the purpose of this research?
The purpose of this research is to evaluate the 'chatbot usability scale'. By doing so, this research might contribute to improving the scale.

Are there possible benefits and risks of participating in this research?
As for benefits, participating might give you more insight into certain methods used during psychological studies. Additionally, you might be able to learn more about chatbots and how to critically view them in the future. Regarding risks, if at any moment you feel uncomfortable during the research, please be reminded that you are free to do or say as pleased. Our study has been reviewed and approved by the BMS Ethics Committee.

What will happen when I want to withdraw from the study?
You can withdraw from the study at any moment if you please. This has no further consequences for you. Moreover, it will still be ensured that all the data collected until that point are deleted and not further used for the study.

Will personal data be collected?
At the beginning, you will be asked some demographical questions (think about age, gender and so forth). This information is important to us to get a complete picture of our participants and to possible gain insight into the effect certain aspects can have on the outcomes of our study. It is your right to request access to and rectification or erasure of personal data.

What will happen with my data?

The collected data will be handled anonymously by removing your name. According to the Netherlands Code of Conduct for Scientific Practice, the data of the study must be stored for at least ten years. This is important to ensure identifiability of the data. In addition, the data might be interesting for further researchers as well and might therefore be confidentiality used in the future.

Contact details

If there are any problems or if you have any questions about the interview, please do not hesitate to contact the researcher:

Amber Ordelman

E-Mail…

Tel.: …

For any other questions or complaints, contact:

Jule Landwehr

E-Mail: ....

Thank you for taking your time to read this information sheet.

Consent Form for *'Artificial Intelligence Conversational Agents: Using Card Sorting To Evaluate The Chatbot Usability Scale'*

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated […-…-22], 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 study involves me filling in a usability scale by myself and undergoing a closed card sorting test. □ □

Use of the information in the study

I understand that information I provide will be used for data analysis and investigating the scale that I am going to fill in during this research. □ □

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 study team. □ □

Future use and reuse of the information by others

I give permission for the data that I provide to be archived in the survey database of the University of Twente so it can be used for future research and learning. My data will be used anonymously as names will be removed and will only be used for research purposes. □ □

Signatures

______________________ ____________________ ________

Name of participant [printed] 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

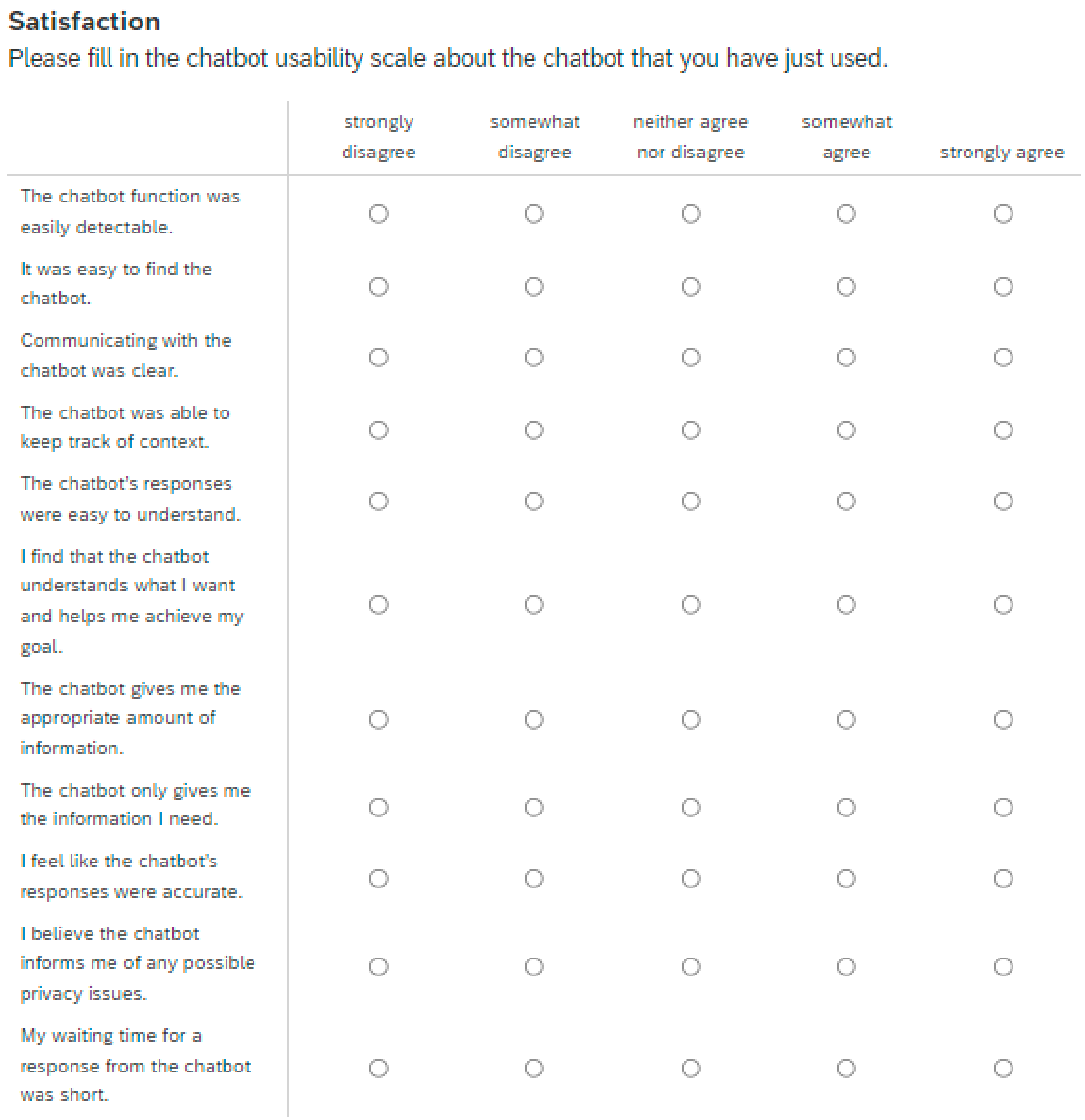
Study contact details for further information:

Amber Ordelman – ....

Contact Information for Questions about Your Rights as a Research Participant
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

the chatbot usability scale (BUS-11) as desgined by Borsci and colleagues



Satisfaction

Please fill in the chatbot usability scale about the chatbot that you have just used.

	strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree
The chatbot function was easily detectable.	○	○	○	○	○
It was easy to find the chatbot.	○	○	○	○	○
Communicating with the chatbot was clear.	○	○	○	○	○
The chatbot was able to keep track of context.	○	○	○	○	○
The chatbot's responses were easy to understand.	○	○	○	○	○
I find that the chatbot understands what I want and helps me achieve my goal.	○	○	○	○	○
The chatbot gives me the appropriate amount of information.	○	○	○	○	○
The chatbot only gives me the information I need.	○	○	○	○	○
I feel like the chatbot's responses were accurate.	○	○	○	○	○
I believe the chatbot informs me of any possible privacy issues.	○	○	○	○	○
My waiting time for a response from the chatbot was short.	○	○	○	○	○

Appendix C

The BUS-11 and its items and factors as originally intended

Factor	Item	
1 - Perceived accessibility to chatbot functions	2	The chatbot function was easily detectable.
	3	It was easy to find the chatbot.
2 - Perceived quality of chatbot functions	11	Communicating with the chatbot was clear.
	12	The chatbot was able to keep track of context.
	13	The chatbot's responses were easy to understand.
3 - Perceived quality of conversation and information provided	14	I find that the chatbot understands what I want and helps me achieve my goal.
	15	The chatbot gives me the appropriate amount of information.
	16	The chatbot only gives me the information I need.
	17	I feel like the chatbot's responses were accurate.
6 - Perceived privacy and security	18	I believe the chatbot informs me of any possible privacy issues.
7 - Time response	12	My waiting time for a response from the chatbot was short.

Appendix D

The trust questionnaire as desgined by McKnight (2011)

	disagree completely (1)	strongly disagree (2)	somewhat disagree (3)	neither agree nor disagree (4)	somewhat agree (5)	strongly agree (6)	agree completely (7)
I experienced chatbots as trustworthy (1)	○	○	○	○	○	○	○
I do not think chatbots will act in a way that is disadvantageous for me (2)	○	○	○	○	○	○	○
Im suspicious of chatbots (3)	○	○	○	○	○	○	○
Chatbot appear deceptive (4)	○	○	○	○	○	○	○
I trust chatbots (5)	○	○	○	○	○	○	○
I experienced to get my question answered when using chatbots (6)	○	○	○	○	○	○	○
Chatbots appear knowledgeable (7)	○	○	○	○	○	○	○
The content of chatbots reflect expertise (8)	○	○	○	○	○	○	○
I feel very confident about the chatbot’s competence (9)	○	○	○	○	○	○	○

Chatbots are well equipped for the task it is set to do (10)	○	○	○	○	○	○	○
Chatbots are natural (11)	○	○	○	○	○	○	○
Chatbots are humanlike (12)	○	○	○	○	○	○	○
Chatbots are realistic (13)	○	○	○	○	○	○	○
Chatbots are present (14)	○	○	○	○	○	○	○
Chatbots are authentic (15)	○	○	○	○	○	○	○
I feel vulnerable when I interact with chatbots (16)	○	○	○	○	○	○	○
I think there could be negative consequences when using chatbots (17)	○	○	○	○	○	○	○
I feel it is unsafe to talk to chatbots (19)	○	○	○	○	○	○	○
I feel I must be cautious when I use chatbots (20)	○	○	○	○	○	○	○
I feel there is risk involved in talking to chatbots (23)	○	○	○	○	○	○	○

Appendix E

All R codes used in the research

R code for the heatmaps:

```
Install.packages(“gplots”)
Install.packages(“RColorBrewer”)
Install.packages(“tidyverse”)
Install.packages(“cluster”)
Install.packages(“factoextra”)
Install.packages(“dendextend”)
Install.packages(“pheatmap”)
```

```
library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

library(RColorBrewer)
library(tidyverse)  # data manipulation

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.8
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse_confli
cts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()

library(cluster)    # clustering algorithms
library(factoextra) # clustering visualization

## Welcome! Want to learn more? See two factoextra-related books at https:/
/goo.gl/ve3WBa

library(dendextend) # for comparing two dendrograms

##
## ---------------------
## Welcome to dendextend version 1.15.2
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
```

```
##
## Suggestions and bug-reports can be submitted at: https://github.com/talg
alili/dendextend/issues
## You may ask questions at stackoverflow, use the r and dendextend tags:
##   https://stackoverflow.com/questions/tagged/dendextend
##
##  To suppress this message use:  suppressPackageStartupMessages(library(d
endextend))
## ---------------------

##
## Attaching package: 'dendextend'

## The following object is masked from 'package:stats':
##
##     cutree
```

```
library(pheatmap)
```

```
example_data <- read.csv("\path\to\data\...\....csv", comment.char="#")
rnames <- example_data[,1]
```

```
mat_data1 <- data.matrix(example_data[,2:ncol(example_data)])
rownames(mat_data1) <- rnames

my_palette <- colorRampPalette(c("yellow","red"))(n = 299)

heatmap.2(dendrogram = "row", mat_data1, col = my_palette, density.info="no
ne", trace="none",
          revC = TRUE, main="Heatmap Example", cexCol = 1, cexRow = 1, marg
ins = c(5, 5))
```

R code for the Kruskal-Wallis test:

```
Levels(extra_data_anova2.csv$trust.group)

library(dplyr) group_by(extra_data_anova2.csv, trust.group)
%>% summarise( count = n(), mean = mean(total.score.correct,
na.rm = TRUE), sd = sd(total.score.correct, na.rm = TRUE),
median = median(total.score.correct, na.rm = TRUE), IQR =
IQR(weight, na.rm = TRUE) )

kruskal.test(total.score.correct ~ trust.group, data =
extra_data_anova2.csv)
```