04/07/2022

Artificial Intelligence Conversational Agents: Using Card Sorting to evaluate the Chatbot Usability Scale (BUS-11), and investigate this in relation to Chatbot Experience

Bachelor Thesis

Lukas Schwemin

S2255995

First supervisor: MSc Jule Landwehr

Second supervisor: Dr. Simone Borsci

University of Twente

BMS Faculty

Department of Psychology

Table of Content

Abstract

Introduction. Chatbot use is rapidly growing worldwide. Especially in the field of customer service, chatbots are regarded as time and cost-effective. As chatbots become more accessible for everyday use, it is important to understand the user's needs in chatbot interactions, to facilitate user uptake. To do so, measuring chatbot satisfaction is the first step. As previous satisfaction measurement tools did not capture the complexity of chatbots, the BUS-11 was developed. To further validate the BUS-11, this study investigated the construct and face validity of the scale. Furthermore, it was tested whether previous experience influences how chatbot satisfaction is perceived.

Methods. Twentythree participants were included in the study. A closed card sorting study was designed to investigate the construct and face validity. Hereby, the construct validity was assessed using heatmaps and the face validity using item-factor tables. Additionally, the participants were grouped based on their chatbot experience level and heatmaps and item-factor tables were plotted for each group. An ANOVA was performed to investigate whether previous experience affects the number of matches with the factorial structure based on the card sorting results.

Results. On average, participants assigned the items to the expected factors during the card sorting. This confirmed the transparency of the construct underlying the scale (face validity). Additionally, the BUS-11 displayed good construct validity due to the participants grouping the items in accordance with the factorial structure. No significant differences were observed between different levels of experience, as the results from each group also mostly confirmed the factorial structure. The ANOVA was not significant but was limited in its statistical power due to the small sample size (the assumption of normality could not be confirmed). The high correlation between factors 2 and 3 found in the original study was also observed here.

Discussion. The results indicate that the BUS-11 provides a reasonable estimate for chatbot satisfaction. It can be said that chatbot experience does not affect the card sorting results and that construct and face validity are good across all chatbot experience levels. The ANOVA had limitations due to sample size but indicated no between-group differences were present.

The participants' mental model seems to fit the factorial structure to a large extent, but it can be suggested that factors 2 and 3 may be combinable.

1. **Introduction**

Chatbot use worldwide is growing rapidly. It is forecasted that the market will reach USD 5.6 billion by 2023, a mean annual increase of almost 100% compared to USD 946 million in 2017 (Businesswire, 2019). Chatbots are defined as any computer or software that can communicate with a human user in natural language (ZEMČÍK, 2019; Io & Lee, 2017). One of the fields chatbots were first used and tested was psychology. In 1966 Joseph Weizenbaum developed a chatbot named ELIZA to replace a psychotherapist. However, the robot was not able to react as a real psychotherapist would and was bound to a given script. Hence, further research was needed. (Weizenbaum, 1983; Io & Lee, 2017). Since then, chatbots have been implemented in other contexts as well. Especially for commercial use and to improve user assistance and guidance, and are now available on many websites and in every new smartphone (e.g., Siri). Especially in the domain of customer services, chatbots are regarded as time and cost-effective (Gnewuch et al., 2017). Therefore, there is a high interest in the development of chatbot usability, as user satisfaction does influence customer loyalty and revenue growth (Gnewuch et al., 2017). As many chatbots on commercial websites did not live up to the expectations and disappeared, there is a need to understand how chatbots can satisfy the customer's needs and design them accordingly (Gnewuch et al., 2017). Additionally, it has been found that even though an increasing number of service providers implement chatbots, the customers' usage lags (Kvale et al., 2021). This supports the need for further research in the domain of chatbot user satisfaction and usability.

As usability limits chatbots' current uptake by users, it is essential to understand and measure this. According to the ISO (2018), usability can be interpreted in terms of user performance and satisfaction. Hereby, user performance is defined by how efficiently and effectively the user reaches the goal, while satisfaction is defined as 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, 2018; ISO, 2019). As user performance can be measured by means of time and accuracy, satisfaction needs to be reflected by the users. Scales measuring satisfaction already exist, and Kvale et al. (2021) used customer satisfaction surveys to measure user experience in chatbots. The study concluded that satisfaction may provide a good reflection of user experience but was not fitting to provide a user experience construct (Kvale et al., 2021). User experience is defined as the “user’s perceptions and responses that result from the use and/or anticipated use of a

system, product or service" (ISO, 2019). To provide a tool capable of measuring the nuances needed to provide a general construct of user experience and satisfaction, van den Bosch and Borsci (2021) argue that chatbots differ from other systems in their diversity and that, therefore, the current scales are not able to provide a sufficient measure of user experience in chatbots.

On this premise, Borsci et al. (2021) designed the Bot Usability Scale (BUS). Initially starting with 42 items, the scale was narrowed down, using factor analysis, to 15 and then to 11 items (Borsci et al., 2021). This is the first questionnaire developed concerning human-computer interaction, especially with regard to artificial intelligence (Borsci et al., 2021). During the study, insights were drawn about the constructs underlying the evaluation of usability in chatbots via factor analysis. During the factor analysis, five factors were identified, measuring chatbot usability (Borsci et al., 2021). While three factors were mutually exclusive, two correlated highly, indicating the need for further research and development (Borsci et al., 2021). Additionally, it was found that the BUS has good reliability and correlates highly with the UMUX-LITE, which indicates good validity in measuring satisfaction (Borsci et al., 2021). Additionally, Borsci et al. (2021) argued that the BUS-15 captures a broader range of chatbot satisfaction aspects, which the UMUX-LITE does not consider. After retesting the BUS-15, the BUS-11 was derived, providing an even better measurement tool for usability in chatbots. The BUS-11 consists of 11 items measuring five factors identified in previous studies: Perceived accessibility to chatbot functions (Items 1 & 2), Perceived quality of chatbot function (Items 3, 4 & 5), Perceived quality of conversation and information provided (Items 6, 7, 8 & 9), Perceived privacy and security (Item 10), and Time response (Item 11) (Appendix A). A card sorting study will be conducted to further investigate and test the factorial structure of the BUS-11 scale.

Card sorting can not only be used to evaluate the usability of website designs but also to investigate humans' mental models. (Olsen-Landis, 2021). Card sorting provides a qualitative approach to assessing mental models in humans. This can be an essential factor in developing measurement tools (like the BUS-11) because it can provide a very detailed description of the mental model of different people, which can be combined to give an estimation of the mental model of the sample. As the BUS-11 was not yet tested using card sorting, it may provide further insights regarding the scale's construct validity and face validity. Face validity is defined as "the extent to which its items subjectively (at first glance) seem to actually cover the constructs they are intended to measure. In the card sort, this was

ascertained by evaluating how many participants had grouped each individual item in the intended construct" (Beerlage-de Jong et al., 2020). Construct validity, on the other hand, is defined as "the extent to which the included items are actually related to each other and the construct they intend to measure". In the card sort, this was evaluated by analysing how the items were grouped together and whether this corresponded with the expected factorial structure (Beerlage-de Jong et al., 2020).

In card sorting, participants sort cards representing one aspect of the topic of interest (e.g. items in a scale) in groups. The results can be utilised to understand which aspects are perceived to be related and belong to the same construct. Hereby there are three card sorting types: open, closed and hybrid (Olsen-Landis, 2021). Open card sorting means that the participant can put the cards together without previously given constructs to sort them in. This method has the benefit that it is very flexible and presents the cognitive model of the participant very accurately, as there are no limitations for the participant on how to sort the cards (Olsen-Landis, 2021). This approach is often used in exploratory studies, as it allows one to understand the mental model underlying the participant's evaluation of the topic of interest (Olsen-Landis, 2021). Closed card sorting is the exact opposite. In close card sorting studies, participants are instructed to sort the card to existing constructs (Olsen-Landis, 2021). This approach is often used to evaluate or test already existing, established models, and confirm their validity (Olsen-Landis, 2021). This approach is less flexible than open card sorting and provides less exploratory results but is, therefore, more controlled and generates more accurate results regarding grouping patterns (Olsen-Landis, 2021). Closed card sorting is less useful in the generative phase of a project (Olsen-Landis, 2021). The third approach to card sorting is the hybrid approach (Olsen-Landis, 2021). Hereby, the participants have given constructs for the card sorting but can add new constructs in case they perceive that a card is not fitting in any of the given constructs (Olsen-Landis, 2021). Therefore, card sorting provides a qualitative assessment of the construct and face validity, adding value to the further development of the scale.

As previous experience was considered in the original study, it is also included in this study (Borsci et al., 2021). Mogaji et al. (2021) found that experience with technology facilitated chatbot use. Additionally, Shih et al. (2006) found a relationship between experience and ease of use in an online classroom. However, experience did not predict the evaluation of the technology (Shih et al., 2006). To investigate this in relation to chatbots, previous chatbot experience will be included as a predictor of card sorting performance.

Hereby, the concept of previous experience consists of familiarity with the concept of chatbots and frequency of use (Borsci et al., 2021). This may provide valuable insights into the user experience construct regarding different levels of experience with chatbots and differences in construct and face validity regarding different levels of experience.

Concluding, this study sets out to answer the questions: (1) Can the face and construct validity of the BUS-11 be confirmed in a card sorting study? (2) Does previous chatbot experience affect the card sorting results, and to what extent do these results match the expected factorial structure? It is hypothesised that (1) the card sorting results will reflect the factorial structure of the BUS-11 and therefore confirm good construct and face validity, (2) that previous chatbot experience affects how the factorial structure is perceived, and that previous chatbot experience has an influence on the extent the card sorting matches the original factorial structure.

2. Methods

2.1 Participants

Twenty-three participants with at least sufficient (intermediate) proficiency in English (B1) participated in the study (13 female, nine male, one unknown, mean age range 17-53). Participants were approached via social media and via the SONA system of the University of Twente. Participants who enrolled via the SONA System were compensated with 0.25 SONA points. Participants under the age of 16 and participants with insufficient English proficiency were excluded. An estimated 40% of the approached participants responded. Before the participants were approached, ethical approval was obtained from the Ethics Committee BMS at the University of Twente.

2.2 Materials

2.2.1 Consent form

The consent form consists of a detailed description of the content and aims of the study, as well as multiple questions regarding taking part in the study, the use of information obtained through the study, and the future use and reuse of the information by others (Appendix B1).

2.2.2 Demographics

The demographic questionnaire includes age, gender identity, nationality and English proficiency (Appendix B2). Regarding gender identity, six options could be chosen, namely: male, female, female-to-male (FTM/Transgender Male/Trans Man), male-to-female (MTF/Transgender Female/Trans Woman), Genderqueer (neither exclusively male nor female), decline to answer, and an additional gender category with the option to specify (in case none of the above was fitting). In the case of nationality, participants could choose between Dutch, German, or other (with the possibility to specify). This was chosen as it was estimated that the majority of participants will be either Dutch or German. Last, the participants could choose four English proficiency levels: B1-Intermediate, B2-Upper Intermediate, C1-Advanced, and C2-Proficient. It was estimated that an intermediate level of English proficiency was sufficient for this study.

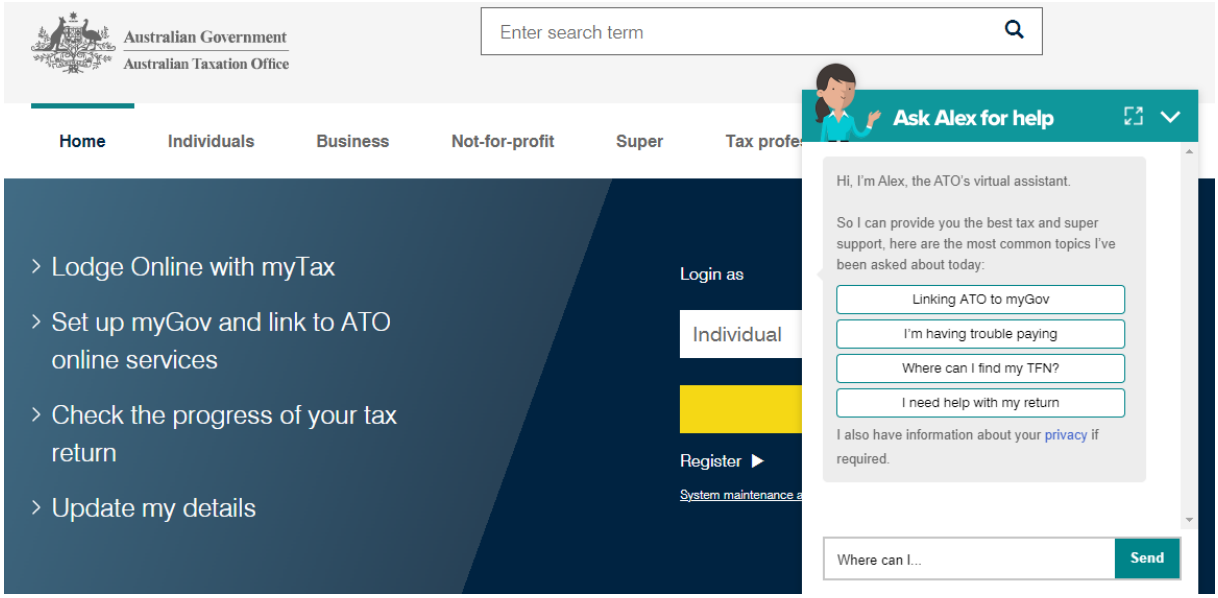
2.2.3 Previous Experience & Trust

The previous experience questionnaire was derived from the original study and consists of 3 items, measured with a 5-point Likert scale (not familiar at all – Extremely familiar; Definitely not – Definitely yes; 0 times – Daily) (Appendix B3) (Borsci et al., 2021). This questionnaire is used to provide an estimation of the experience with chatbots.

Regarding the trust questionnaire, items from McKnight were adapted to the most relevant constructs of technology regarding trust. The items are scored on a 5-point Likert scale. This questionnaire will not be taken into account in this report and therefore is not relevant for this study (it is relevant for another study, which is based on the same questionnaire and participants).

2.2.4 Chatbots

Two chatbots were used to give the participants an understanding of the topic of the study (Appendix B4). The chatbots were derived from the original study. The first chatbot is A&O Hostels, a website for booking, comparing and evaluating hostel rooms (Appendix B4.2). The second chatbot is Alex and can be found on the website of the Australian Taxation Office (Appendix B4.3).



2.2.5 BUS 11-item-scale

The Bot Usability Scale (BUS) consists of 11 items measured on a 5-point Likert scale (strongly disagree – strongly agree) and is the original questionnaire, of which the validity will be tested in this study. It measures user satisfaction and usability regarding the use of chatbots.

2.2.6 Data analysis

The software Qualtrics and R were used for the data collection and analysis, respectively.

2.3 Procedure

The Participants were informed about the needed English proficiency level, that the survey takes approximately 15 minutes and that it is about Usability in Chatbots. After that, the participant could decide whether to participate or not. If they chose to participate, they had to give informed consent at the beginning and fill in the survey without supervision. The survey was administered using Qualtrics.

After the informed consent was obtained, the demographics and the level of previous experience with chatbots were asked. This was followed by the tasks to be completed with the two chatbots, with the addition of the BUS-11 and Trust questionnaires. At the end of the survey, the participants were asked to participate in card sorting regarding the items of the BUS-11 they had previously filled in (Appendix B5). After that, the survey ends.

2.4 Data Analysis

To answer the first research question and test the construct and face validity of the BUS-11, the card sorting results were plotted in a Heatmap (item-item matrix) and an item-factor table. To plot the heatmap, the card sorting results were first transcribed into Jaccard coefficients and plotted in a similarity matrix.

2.4.1 Item-Item matrix for construct validity

2.4.1.1 Jaccard coefficient (similarity measure)

The collected data will be analysed using the Jaccard coefficient and displayed in a similarity matrix. The Jaccard coefficient is used to assess the similarity between two groups on a scale of 0% to 100%. In card sorting, this is used to analyse which cards were sorted together, to later see what items form clusters together. These clusters are then compared to the item groupings in the factorial structure. Based on this, it can be assessed to what extent the expected factorial structure of the BUS 11 is replicated in the card sorting, providing insights regarding the construct validity. The Jaccard coefficient is calculated in two steps: (1) recording the number of groups both items were grouped in (number A) and recording the number of groups either of the two items were grouped in (number B); (2) dividing number A by number B (dividing the number of groups both items were grouped in, by the number of groups either of the two items were grouped in) (Schmettow & Sommer, 2016). This procedure is done for every participant individually. After recording all participant's answers, the answers are grouped in one similarity matrix.

2.4.1.2 Creating the Heatmap

After all the Jaccard scores are recorded and combined in one matrix, this matrix will be reordered and colour-coded to make the similarity groups visible and investigate the scale's construct validity (Appendix A). The reordering will include deleting missing values, transforming the data set into a numeric format and giving names to the items, as well as changing the order of the items so items with high levels of agreement (frequently sorted together) are grouped together. The colour coding will be yellow for low agreement levels and red for high agreement levels (typical colours for a heat map).

2.4.2 Item-Factor table for construct and face validity

To assess the face validity of the scale, it will be recorded how frequently the items were grouped in the factors. The technique will be closely connected to the Jaccard coefficient. First, how many times every item was grouped into every category will be

counted. Second, this number will be divided by the overall responses (the frequency, the item could have been grouped in the categories if every participant grouped the item in the same group). The results are a percentage of how frequently every item was grouped in every category. These will be displayed in a matrix, where it should become apparent that items are grouped more frequently in specific groups. Based on this, a conclusion can be drawn on the face validity of every item and category (and the overall scale). Additionally, construct and face validity are assessed by comparing the groupings to the expected factorial structure of the BUS-11.

2.4.3 Chatbot Experience Measure

To answer the second research question and test whether differences in experience with chatbots influence the card sorting results, the participants will be grouped according to their experience level. Hereby, three groups were created based on their mean experience levels. The group with the lowest level consists of 10 participants (Group 1- Low Experience), the second group with medium experience consists of 6 participants (Group 2 – Medium Experience) and the third group with high experience of 7 participants (Group 3 – High Experience). An individual heatmap and item-factor table will be created for every group. Based on this, it was analysed whether the groups sorted the cards differently, which can affect the construct validity, as well as the face validity of the scale (e.g., maybe it measures user satisfaction in chatbots more accurately depending on your level of experience).

2.4.4 Investigating group differences

An ANOVA will be performed to answer the second part of the second research question and assess whether previous experience affects the number of matches of the card sorting results with the original factorial structure. Hereby, the same groups will be used. First, a Boxplot analysis will be performed to observe whether differences between groups are visible. Second, the ANOVA will be performed to investigate the group differences further. It was assessed whether groups differ in the percentage of matches with the original factorial structure. Additionally, to ensure the validity of the ANOVA, the assumption of Normality, Homogeny of Variance, and Independence will be tested using a Histogram, together with the Shapiro-Wilk test (Normality) (Appendix D1) and the Levene test (Homogeneity of Variance) (Appendix D2), respectively. The assumption of Independence is

given, due to the between-subject design and the sampling procedure, as the groups did not influence each other.

3. Results

3.1 Construct and Face validity of the BUS-11

A heatmap and an item-factor table were conducted to answer the first research question and test the construct and face validity of the BUS-11. Hereby, the heatmap can give insights regarding the construct validity and the item-factor table regarding the face and construct validity of the scale.

3.1.1 Item-to-item Heatmap for construct validity

A heatmap was created based on the card sorting results to investigate whether participants grouped the items according to the factorial structure and therefore test the scale's construct validity. Hereby it becomes apparent which items were sorted together frequently. This becomes visible after colour coding. The colour scale ranges from yellow (not sorted together) to red (always sorted together). Hereby the colours correspond with a number of 0 to 23, as no participant can group the items together (0%) or all participants group the items together (100%) and everything in between. Utilising this technique, conclusions about the construct validity can be drawn.

As displayed in Figure 1 by the red coloured squares in the top left corner, participants agree that Items 10 and 11 should be independent factors, as they display almost no mutual sorting with other items. This corresponds with the original study, where items 10 and 11 measure factors 4 and 5, respectively. Additionally, it can be seen that items 1 and 2 were very often grouped together. However, not frequently with other items, which indicates that participants perceive them to measure one factor themselves, which is in accordance with the original study, where items 1 and 2 fall together in factor 1.

Looking at the other items, two groups become apparent. The first group, which is relatively closely clustered, consists of items 3, 5, 7 and 9. Especially items 3 and 5 have a high level of agreement, as well as items 5 and 7. Regarding item 9, participants frequently sorted it with the other three, but the agreement seems not as strong compared to the other relations (between 3, 5 and 7). The second group consists of items 4, 6 and 8. Especially regarding items 4 and 6, participants highly agree with each other. Item 8, on the other hand,

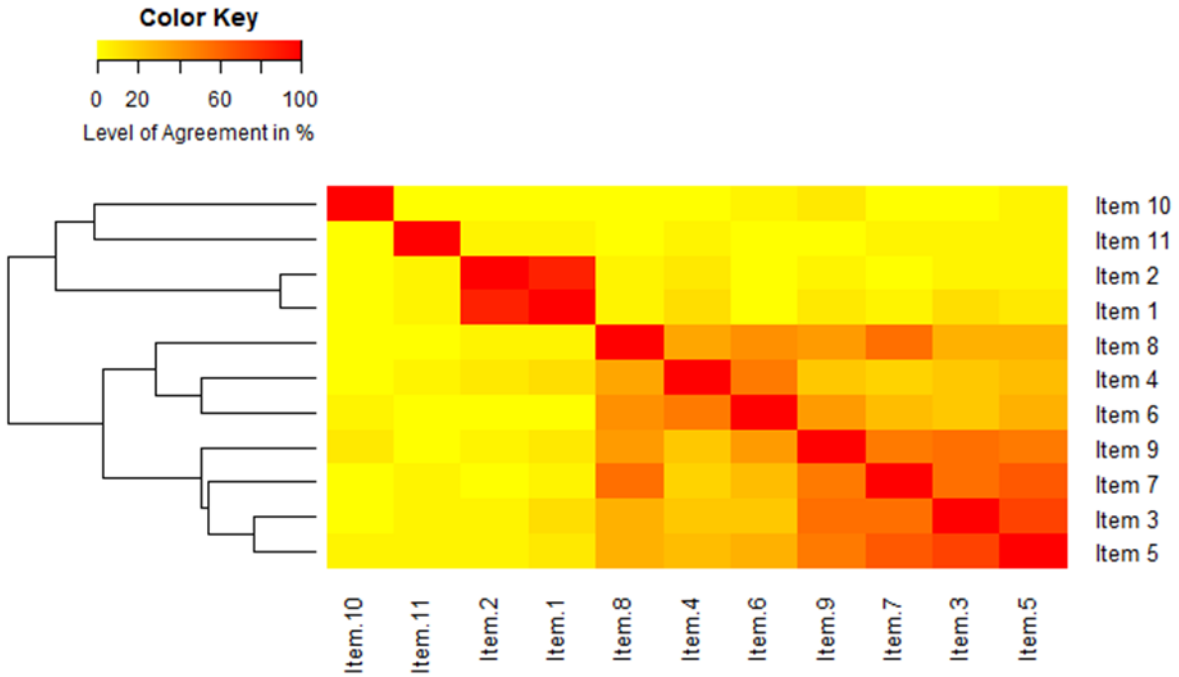
has a moderate to high level of agreement with items 4 and 6 but was also frequently grouped with item 7, as well as moderately frequent with items 3, 5 and 9. As a sidenote, participants also show a moderately strong agreement regarding items 6 and 9.

These findings differ from the original study, where items 3, 4 and 5 fall into one factor, as well as items 6, 7, 8 and 9. This means that they should have a high level of agreement if they fall into the same factor. The card sorting revealed that items 3 and 5 were not frequently grouped with item 4 but with items 7 and 9. Additionally, item 6 was frequently grouped with item 4 but not as frequently with items 7 and 8.

In conclusion, these results display a high similarity with the original factorial structure, especially regarding factors 1 (items 1 & 2), 4 (item 10) and 5 (item 11). The factorial structure regarding factors 2 and 3 differs slightly from the original but includes the same items. Overall, the scale displayed good construct validity.

Figure 1.

Heatmap of the frequency of item groupings (including all 23 participants). The colour key ranges from yellow (0% of participants grouped the items together) to red (100% of participants grouped the items together). The items are rearranged according to their agreement level with other items to make clusters clearly visible.



3.1.2 Item to factor table for face and construct validity

The Item-to-factor table was conducted to test the face and construct validity of the scale. Hereby it can be observed how frequently certain items were sorted to certain factors (face validity) and whether the sorting matched with the factorial structure (construct validity).

It is shown in Table 1 that 91% of the participants grouped items 1 and 2 in factor 1, as well as item 10 in factor 4, and even 96% put item 11 in factor 5. These results are in accordance with the original factorial structure. Regarding factors 2 and 3, the sorting frequencies were more distributed over the factors. Table 1. shows, that participants sorted item 4 (61%) and 6 (70%) in factor 2 and items 3 (61%), 5 (74%), 7 (87%), 8 (57%) and 9 (61%) in factor 3. In contrast, in the original factorial structure, factor 2 consists of items 3, 4 and 5 and factor 3 of items 6, 7, 8 and 9. Additionally, it becomes apparent that item 8 has the most variance regarding the sorting frequencies of participants because it was sorted to factors 1-3, with a very similar value displayed in Table 1, regarding factors 2 and 3. The higher percentage regarding factor 3 is similar to the original factorial structure. Despite item 8 being the item varying the most across factors, the other items belonging to factors 2 and 3 in the original factorial structure also display a more equal distribution across factors (compared with items 1, 2, 10 and 11) and are therefore also only partially matching the results from the original study. Contrary to that, item 7 is the only one that originally belonged to factor 3, which was also sorted with a high frequency in said factor. Therefore, item 7 exceeds the 75% agreement level threshold and matches the original factorial structure.

In conclusion, the results displayed in Table 1 are congruent with the factorial structure in the original study, and therefore good construct and face validity are displayed. Even though the majority of the items only match the original partially (the items belonging to factors 2 and 3, apart from item 7), there is a match of all items to some degree. Especially, items 1, 2, 10 and 11 matched the original factorial structure, with over 90% of participants sorting them in accordance with the existing model.

Table 1.

General item-level agreement matrix. The table displays the frequency (in % of participants) of items being sorted to a factor (face validity). Hereby it is also displayed whether the

original factorial structure was matched, partially matched or not matched at all (construct validity). The threshold for a match is 75% agreement.

Original factor	**Item**	ACCES[a](%)	QUAL[b](%)	CONV[c](%)	PRIV[d](%)	TIME[e](%)	**Match with original factorial structure**
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	PARTIAL
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 (Factor 1)

QUAL[b] = Perceived quality of chatbot functions (Factor 2)

CONV[c] = Perceived quality of conversation and information provided (Factor 3)

PRIV[d] = Perceived privacy and security (Factor 4)

TIME[e] = Time response (Factor 5)

3.2 Chatbot Experience and Card Sorting: a qualitative observation

To answer the second research question and assess whether different levels of experience with chatbots affect the card sorting results, heatmaps and item-factor tables were made for each group/level of experience. Based on these, differences and similarities compared to the original study and across groups are described. Additionally, this gives insights about construct and face validity regarding different levels of experience.

3.2.1 Low Level of Chatbot Experience

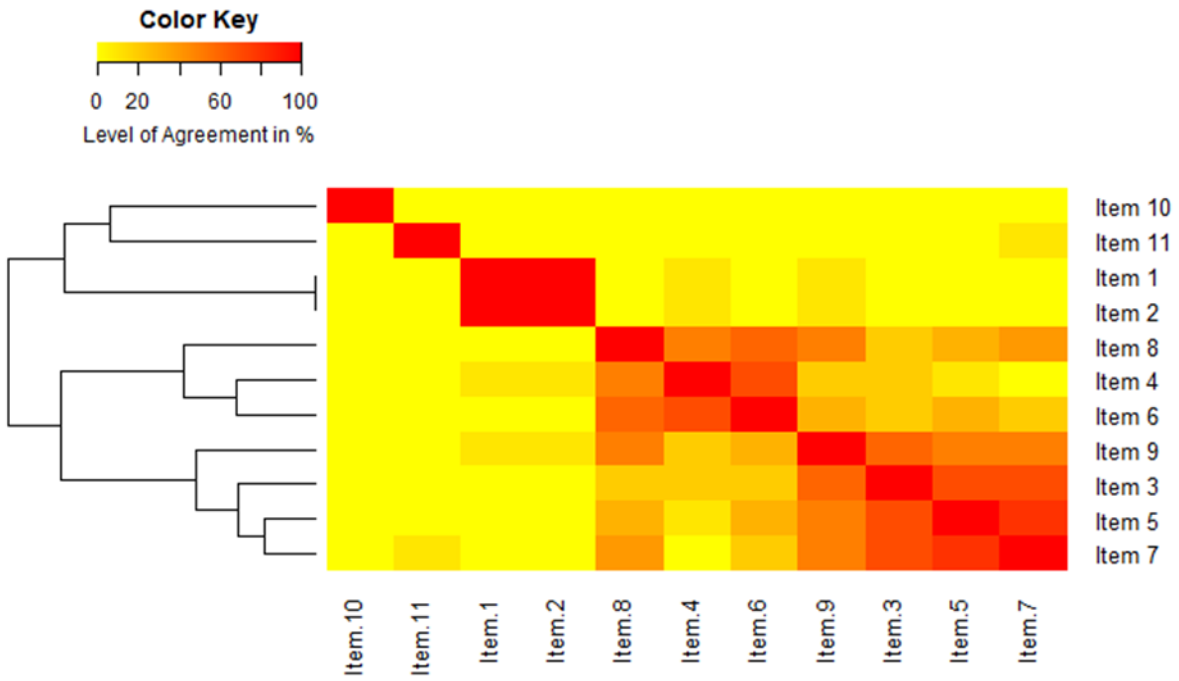
To draw conclusions about the construct validity in regard to low experience levels, the Heatmap in Figure 2 was conducted. Figure 2 displays a cluster consisting of items 1 and 2, as well as individual clusters consisting of item 10 and item 11 (top left corner of the heatmap). There is a strong similarity to the original study, which also reported these clusters after factor analysis. The other seven items are clustered in a more complex pattern. Even though there is a lot of agreement across participants regarding all of those seven items, two cluster-like structures can be observed in the heatmap. The first consists of items 4, 6 and 8, and the second one of items 3, 5, 7 and 9. This is similar to Figure 1, as the same clusters are visible in the heatmap, even though the colour coding does not as strongly indicate them.

The distribution of item 8, observed in Figure 1, can also be observed here. It can be seen in Figure 2 that item 8 was moderately frequently grouped with items 7, 9, 6 and 4 and less frequently with item 5. Compared to the original study, this is similar, as item 8 belongs to the same factor as items 6,7 and 9 in the original factorial structure. On the other hand, the frequent groupings with items 4 and 5 stand in contrast to the existing model, as they belong to a different factor. Additionally, item 8 was frequently grouped with items from both clusters, observable in Figure 2 (centre to bottom right).

In conclusion, the original factorial structure was, to a large extent reproduced for this level of chatbot experience, showing a sufficient level of construct validity. Especially regarding items 1, 2, 10 and 11, there was a high level of agreement amongst participants. Even though this level of agreement was not so high for the other items, two additional clusters can still be observed among those items (Figure 2). Item 8 was frequently grouped with items from both of these clusters and items belonging to factors 2 or 3 in the original factorial structure.

Figure 2.

Heatmap of the frequency of item groupings (including 10 participants with a low level of chatbot experience – Group 1). The colour key ranges from yellow (0% of participants grouped the items together) to red (100% of participants grouped the items together). The items are rearranged according to their agreement level with other items to make clusters clearly visible.



For investigating the face validity of the BUS-11 in regard to low chatbot experience, the frequencies of items being sorted to factors are displayed in Table 2. Additionally, the matches with the original factorial structure provide insights into the scale's construct validity. It can be observed in Table 2 that items 1 and 2 were indeed sorted by all participants in factor 1, similar to the original factorial structure. This also applies to item 10 and item 11, sorted by 100% of the participants in factors 4 and 5, respectively. Also, similar to the original study, no other items were sorted in these factors by a substantial proportion of the group (not higher than 10% of participants). Items 4 and 6 were sorted with a high frequency in factor 2. Hereby, item 4 was sorted similar to the original study, while item 6 was not. Item 6 was more frequently grouped with item 4 than with the items belonging to factor 3 (items 7, 8 and 9), to which item 6 also belongs in the original model. Items 3, 5 and 7 were sorted in factor 3 with a high frequency. As items 3 and 5 originally belong in factor 2, this differs from the original study. Items 8 and 9 were sorted in factors 2 and 3, with equal distributions across the two factors. This supports the high correlation between these factors found in the original study. Apart from items 8 and 9, only items 3 and 6 partially matched the original factorial structure, while the other items fully matched it.

In conclusion, the factorial structure regarding factors 2 and 3 in Table 2 differs from the original, while factors 1, 4 and 5 are similar to the original, displaying good face validity.

Additionally, the items sorted to factors 2 and 3 are similar to the original factorial structure but are differently distributed across these factors. Overall, the original factorial structure was reproduced, as only items 3, 6, 8 and 9 matched the original model partially, while the other items fully matched the model after applying the threshold of 75%. Therefore, good construct validity regarding low levels of chatbot experience can be concluded.

Table 2.

Group 1 - Item-level agreement matrix of 10 participants with low chatbot experience level. The table displays the frequency (in % of participants) of items being sorted to a factor (face validity). Hereby it is also displayed whether the original factorial structure was matched, partially matched or not matched at all (construct validity). The threshold for a match is 75% agreement.

Original factor	**Item**	ACCES[a](%)	QUAL[b](%)	CONV[c](%)	PRIV[d](%)	TIME[e](%)	**Match with original factorial structure**
ACCES	1	100	0	0	0	0	YES
ACCES	2	100	0	0	0	0	YES
QUAL	3	0	30	70	0	0	PARTIAL
QUAL	4	10	80	10	0	0	YES
QUAL	5	0	20	80	0	0	YES
CONV	6	0	70	30	0	0	PARTIAL
CONV	7	0	10	80	0	10	YES
CONV	8	0	50	50	0	0	PARTIAL
CONV	9	10	40	50	0	0	PARTIAL
PRIV	10	0	0	0	100	0	YES
TIME	11	0	0	0	0	100	YES

ACCES[a] = Perceived accessibility to chatbot functions (Factor 1)

QUAL[b] = Perceived quality of chatbot functions (Factor 2)

CONV[c] = Perceived quality of conversation and information provided (Factor 3)

PRIV[d] = Perceived privacy and security (Factor 4)

TIME[e] = Time response (Factor 5)

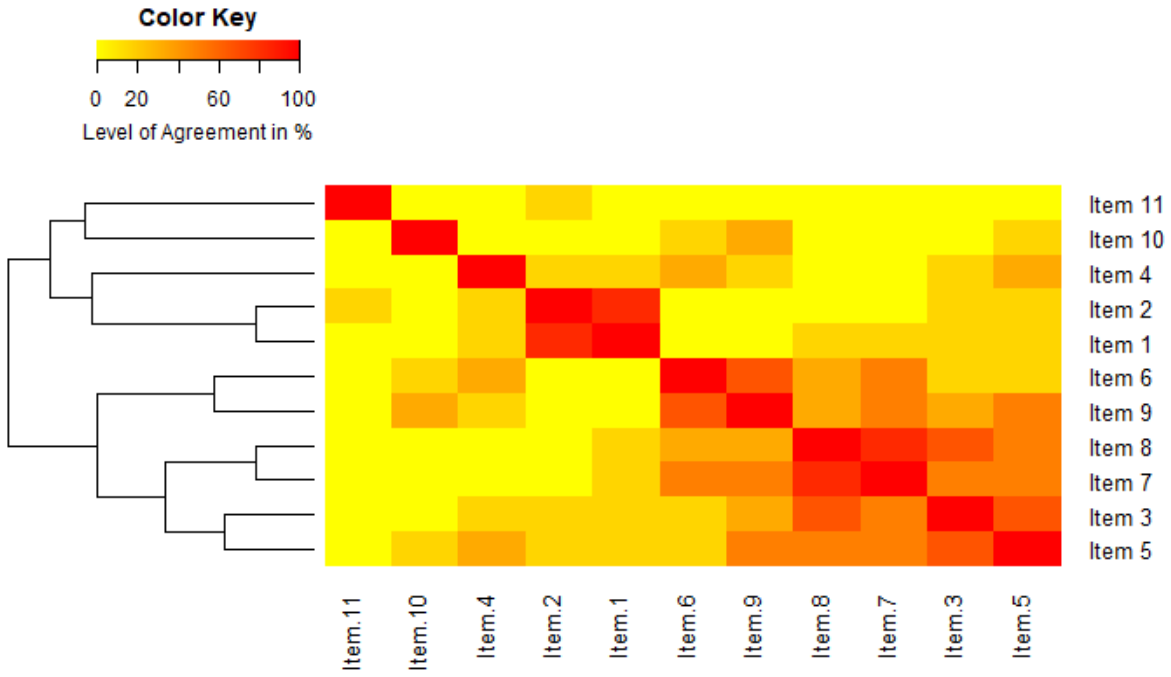
3.2.2 Medium Level of Chatbot Experience

To investigate the construct validity regarding medium levels of chatbot experience, the heatmap in Figure 2 was plotted. The Heatmap shows, similar to the low chatbot experience group that items 1 and 2 were clustered together, without frequent groupings with other items. Regarding items 4, 10, and 11, participants did not group them frequently with other items (item 4 was sometimes grouped with item 6, which was also displayed in regard to low chatbot experience (Group1)), in contrast to group 1 and the original study, where this only applied to items 10 and 11 and not to item 4. There is even more distribution for the other six items than in group 1. Three clusters can be observed consisting of items 3, 5, 6, 7, 8 and 9. Even though there are moderately strong agreement levels regarding all of these items (apart from item 6, which was not frequently grouped with items 3 and 5), it can be observed that items 6 and 9 were grouped frequently together, as well as items 7 and 8, and items 3 and 5. Apart from the strong agreement regarding items 6 and 9, the other between-item agreements were also displayed in group 1. Even the agreement regarding items 6 and 9 was similar to the original study. In contrast to group 1, regarding items 3 and 9, participants only had a weak level of agreement. Additionally, according to this heatmap, there are 6 or 7 factors underlying the items, which strongly contrasts the original factorial structure with five factors.

In conclusion, contrary to the low experience group, the whole sample and the original factorial structure, participants in this group agree on item 4, referring to an individual factor, as it was not frequently grouped with other items. Additionally, the items originally belonging to factors 2 and 3 (excluding item 4) displayed a wide distribution regarding the levels of agreement. Even though the factorial structure of the original study was largely reproduced, as items 1 and 2 were clustered together, as well as items 10 and 11, who make up clusters by themselves. These results are in accordance with the results from the original study. Therefore, the construct validity was lower than in the low chatbot experience group but still sufficient, as item 4 did not match the original factorial structure.

Figure 3.

Heatmap of the frequency of item groupings (including 6 participants with a medium level of chatbot experience – Group 2). The colour key ranges from yellow (0% of participants grouped the items together) to red (100% of participants grouped the items together). The items are rearranged according to their agreement level with other items to make clusters more clearly visible.



With the purpose of investigating the face and construct validity regarding medium levels of chatbot experience, the frequency of items being sorted to factors and whether they match the factorial structure was assessed and is displayed in Table 3. The wide distribution of agreement levels in this group can be observed displayed in Table 3, as 7 of the 11 items only partially match the original factorial structure. Additionally, almost all items were sorted into two or more factors. Similar to group 1 and the original study, items 1 and 2 were grouped in factor 1 with a high frequency, as well as item 11 in factor 5. Item 10, which was frequently grouped in factor 4 in the low chatbot experience group and the overall sample, was not as frequently grouped in factor 4 in the medium chatbot experience group. Here only 67% grouped item 10 in factor 4, and therefore the agreement levels regarding item 10 do not exceed the threshold of 75%, and it only partially matches the original model (this can also be attributed to the small sample size in this group). This is a big difference compared to the 100% in the low experience group, as item 10 was grouped in two other factors. On the other hand, similar to the group with low chatbot experience, items 4 and 6 were grouped with a relatively high frequency in factor 2, even though item 4 was sorted in 4 different factors and

partially matched the original factorial structure. Items 3, 5, 7 and 9 were sorted frequently in factor 3. Item 8 was, similar to the low experience group and the overall sample, distributed across factors, with (like item 4) the highest agreement level of 50%. Similar to the low chatbot experience group, the overall sample and the results from the original study, items 1, 2, and 11, matched the original factorial structure. Additionally, contrary to the group with low experience and the overall sample, item 5 matched the original model. The other items only partially matched the factorial structure of the BUS-11.

In conclusion, the items were more scattered across the factors compared to the low experience group but also showed similarities regarding the matches with the factorial structure. Especially regarding factors 1 and 2, the factorial structure was reproduced. Additionally, the distribution of agreement levels across factors 2 and 3 was also similar to the low experience group and the overall sample. Contrary to these two groups, item 10 only partially matched the factorial structure and was also sorted to other factors. Overall, good face and construct validity could be confirmed due to the matches with the original factorial structure, even though there was less agreement compared to the low experience group and the overall sample.

Table 3.

Group 2 - Item-level agreement matrix of 6 participants with medium chatbot experience level. The table displays the frequency (in % of participants) of items being sorted to a factor (face validity). Hereby it is also displayed whether the original factorial structure was matched, partially matched or not matched at all (construct validity). The threshold for a match is 75% agreement. Values are rounded to two decimal places.

Original factor	**Item**	ACCES[a](%)	QUAL[b](%)	CONV[c](%)	PRIV[d](%)	TIME[e](%)	**Match with original factorial structure**
ACCES	1	100	0	0	0	0	YES
ACCES	2	83.33	0	0	0	0	YES
QUAL	3	16.67	16.67	66.67	0	0	PARTIAL
QUAL	4	16.67	50	16.67	16.67	0	PARTIAL
QUAL	5	16.67	0	83.33	0	0	YES

CONV	6	0	66.67	33.33	0	0	PARTIAL
CONV	7	16.67	16.67	66.67	0	0	PARTIAL
CONV	8	16.67	33.33	50	0	0	PARTIAL
CONV	9	0	33.33	66.67	0	0	PARTIAL
PRIV	10	0	16.67	16.67	66.67	0	PARTIAL
TIME	11	0	0	0	0	100	YES

ACCES[a] = Perceived accessibility to chatbot functions (Factor 1)

QUAL[b] = Perceived quality of chatbot functions (Factor 2)

CONV[c] = Perceived quality of conversation and information provided (Factor 3)

PRIV[d] = Perceived privacy and security (Factor 4)

TIME[e] = Time response (Factor 5)

3.2.3 High Level of Chatbot Experience

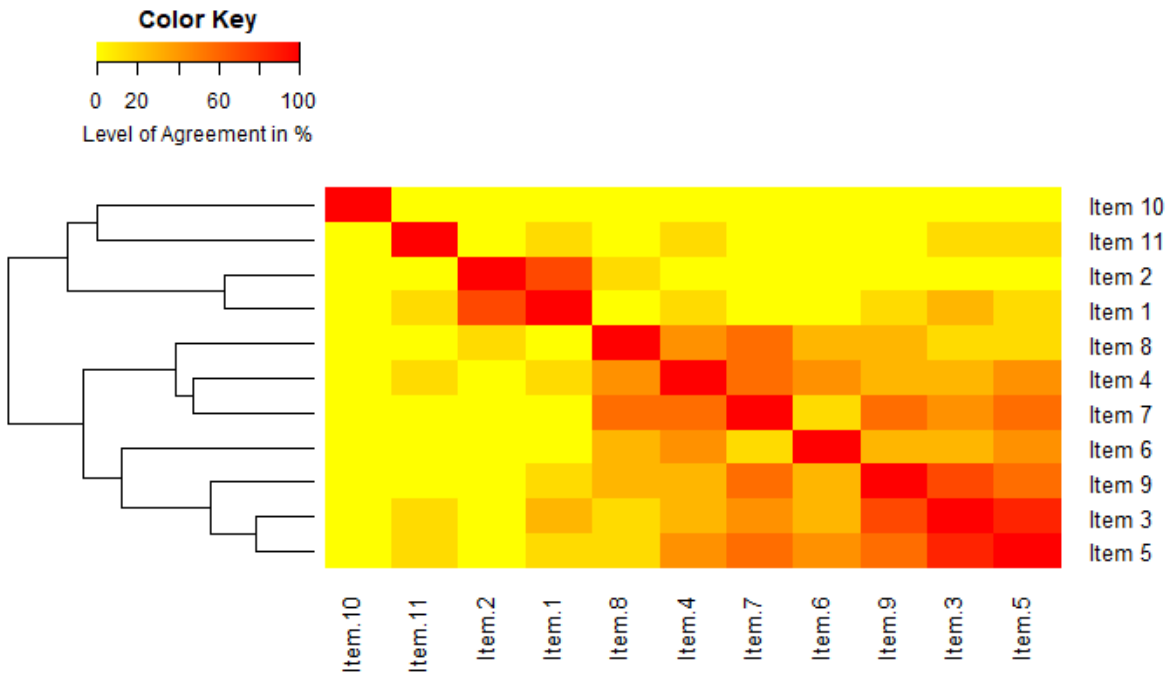
To investigate the construct validity in regard to high chatbot experience, the heatmap in Figure 4 was plotted. Like in the groups with low and medium chatbot experience levels, the heatmap displays that items 1 and 2 were clustered together without being grouped with significant frequencies with other items, as well as that the participants perceive items 10 and 11 to refer to an individual factor. Additionally, two clusters become apparent. The first one includes items 3, 5 and 9, which were frequently grouped together, especially items 3 and 5. Item 9 was frequently grouped with item 3 but less frequently with item 5. This shows a contrast to the original study, where item 9 is related to a different factor than items 3 and 5, but also displays similarity to the other groups, as there was a high level of agreement regarding those items. Additionally, the high agreement level regarding items 3 and 5 is similar to the original factorial structure. The second cluster consists of items 4, 7 and 8, with item 6 frequently being grouped with items 4 and moderately frequent with item 8 but not at all with item 7. This differs significantly from the original study, where items 6 and 7 belong to the same factor. Additionally, item 4 belongs to a different construct than items 6, 7 and 8 in the original factorial structure but was grouped frequently with all three items. This also displays a similarity to the low experience group, where item 4 was also frequently sorted with item 6. It can be said that item 6 differs from the original factorial structure due to the

high agreement levels regarding being grouped with items 4 and 5 and not being grouped with item 7.

In conclusion, sufficient construct validity was displayed, as the card sorting results match the factorial structure to a large extent, especially regarding items 1, 2, 10 and 11. The clusters observed in the two other groups, especially the low experience group, were also displayed here. On the other hand, item 6 was not frequently grouped with other items and was not grouped with item 7, which stands in contrast to the factorial structure, where items 6 and 7 belong in the same factor and therefore highly correlate. The wide distribution regarding the agreement levels of item 8 was not as present as in the previous groups but could still be observed to some degree.

Figure 4.

Heatmap of the frequency of item groupings (including 7 participants with a high level of chatbot experience – Group 3). The colour key ranges from yellow (0 or 0% of participants grouped the items together) to red (7 or 100% of participants grouped the items together). The items are rearranged according to their agreement level with other items to make clusters more clearly visible.



For assessing the face and construct validity regarding high levels of chatbot experience, the frequencies of items being sorted to factors and whether they match the

factorial structure were plotted in Table 4. Again, like the group with low and medium experience levels, items 1 and 2 were sorted frequently in factor 1. While the agreement level regarding item 2 was around 85%, the agreement level regarding item 1 was only around 70%, not exceeding the threshold of 75%. Therefore, item 1 only partially matches the factorial structure, which contrasts the previous findings in the other groups. Additionally, all participants sorted item 10 in factor 4 and 85% sorted item 11 in factor 5. Therefore items 2, 10 and 11 fully match the factorial structure. The items originally belonging to factors 2 and 3 are mostly distributed across the two factors, apart from item 7, which was sorted by 100% of participants in factor 3. Item 7 matches the original factorial structure as well. Additionally, items 3, 4 and 5 are almost equally distributed across factors 2 and 3. This is in accordance with the results from the other groups and the correlation regarding these factors described in the original study. Like in the groups with low and medium chatbot experience, item 6 was frequently sorted in factor 2 (by 70% of participants), which differs slightly from the original study, as item 6 also partially matches the factorial structure. Item 3 was sorted by around 57% in factor 2, partially matching the factorial structure. This contrasts the other groups, where the majority (over 50%) sorted item 3 in factor 3. Additionally, items 8 and 9 were sorted in factor 3 and partially match the factorial structure, as the agreement level does not exceed the threshold of 75%. Item 4 was sorted in factor 3, contrary to the other groups' sorting and the original study. The correlation regarding factors 2 and 3 from the results of the original study can also be observed here, as the items belonging to said factors were almost equally distributed across these factors.

In conclusion, good face and construct validity were confirmed, as the factorial structure was reproduced to a large extent, especially regarding items 2, 10 and 11. Contrary to the other groups and the overall sample, item 1 partially matched the factorial structure. The items belonging to factors 2 and 3 were distributed across both factors (as well as across others for some items), which matches the results from the other groups as well as the overall sample and also represents the correlation between factors 2 and 3 described in the original study. Item 7 must be excluded in this regard, as it fully matched the factorial structure, with all participants sorting it to factor 3.

Table 4.

Group 3 - Item-level agreement matrix of 7 participants with high chatbot experience level. The table displays the frequency (in % of participants) of items being sorted to a factor (face validity). Hereby it is also displayed whether the original factorial structure was matched, partially matched or not matched at all (construct validity). The threshold for a match is 75% agreement. Values are rounded to two decimal places.

Original factor	**Item**	ACCES[a](%)	QUAL[b](%)	CONV[c](%)	PRIV[d](%)	TIME[e](%)	**Match with original factorial structure**
ACCES	1	71.43	28.57	0	0	0	PARTIAL
ACCES	2	85.71	0	0	0	14.29	YES
QUAL	3	0	57.14	42.86	0	0	PARTIAL
QUAL	4	0	42.86	57.14	0	0	PARTIAL
QUAL	5	0	42.86	57.14	0	0	PARTIAL
CONV	6	14.29	71.43	14.29	0	0	PARTIAL
CONV	7	0	0	100	0	0	YES
CONV	8	14.29	28.57	57.14	0	0	PARTIAL
CONV	9	14.29	28.57	57.14	0	0	PARTIAL
PRIV	10	0	0	0	100	0	YES
TIME	11	0	14.29	0	0	85.71	YES

ACCES[a] = Perceived accessibility to chatbot functions (Factor 1)

QUAL[b] = Perceived quality of chatbot functions (Factor 2)

CONV[c] = Perceived quality of conversation and information provided (Factor 3)

PRIV[d] = Perceived privacy and security (Factor 4)

TIME[e] = Time response (Factor 5)

even though visual differences between Figures 1, 2, 3 and 4 were observed, the expected factorial structure was at least partially confirmed in all four heatmaps. Apparently, when it comes to participants' mental model, items 1, 2, 10 and 11 are placed in full accordance with the original factorial structure and sorted in factors 1, 4, and 5, respectively

(items 1 and 2 both belong to factor 1). The items belonging to factors 2 and 3 were more distributed across the two factors in line with the correlation of these factors identified in the original study.

Additionally, a few items (especially item 8) displayed almost the same frequency regarding being sorted in factors 2 and 3. These results were mostly constant over the groups, and only differences regarding single items were observed. Overall, construct and face validity were confirmed regarding the three levels of chatbot experience, as the factorial structure was visible in all three groups.

3.3 Effects of experience with chatbots on the card sorting

An ANOVA analysis with the three groups described above will be performed to further answer the second research question and statistically assess whether previous chatbot experience affects the percentage of matches with the factorial structure from the original study. Herby, it will be assessed whether there are significant differences between groups of different levels of experience regarding the match of their mental model with the factorial structure. For the visualisation of the group differences, a Boxplot was plotted.

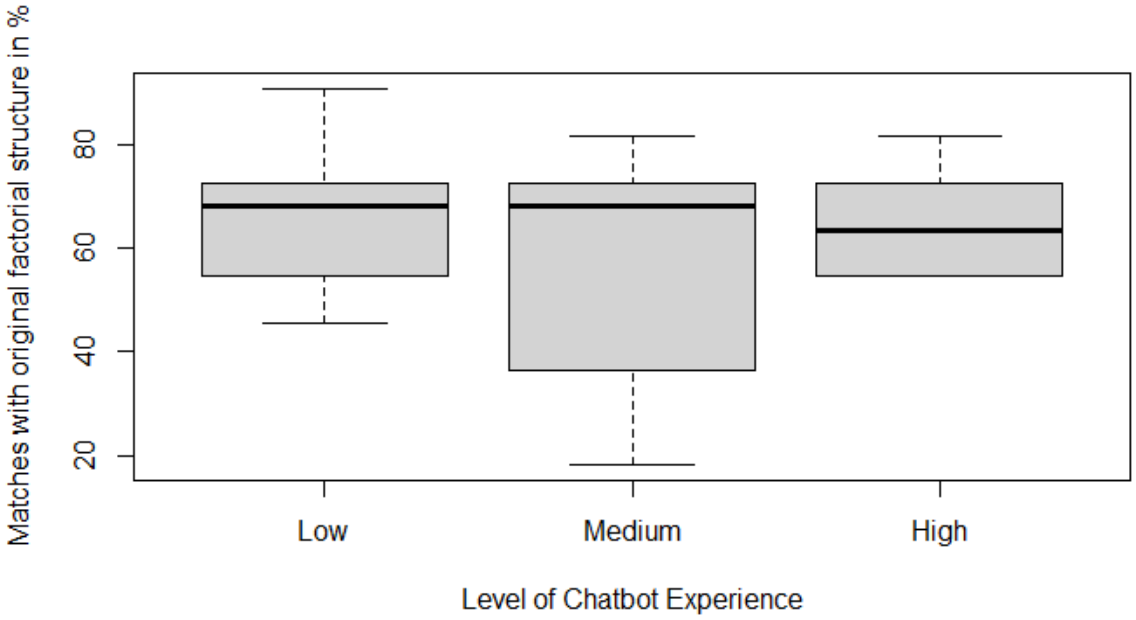
(Figure 6). No significant differences across the mean group levels can be observed. It becomes apparent that the group with a medium level of chatbot experience has more low scores than the group with low experience, even though the medium experience group has a higher experience level. Additionally, the distribution of groups with low and high chatbot experience seems quite similar, with the high chatbot experience group even displaying a lower mean than the group with low and medium experience. Additionally, it can be observed that the high experience group has no outliers in the lower spectrum, indicating a tendency for high scores, which stands contrary to the lower mean (compared to the other groups). The scores of the group with high chatbot experience are more closely clustered, while the medium experience group's scores are more distributed over the scale of the matches with the factorial structure. On the other hand, the group with low experience displays high scores, containing the outlier with the highest score, but is also clustered together, like the high experience group.

In conclusion, it can be said that the boxplot did not display any significant group differences. The group with medium chatbot experience was more distributed across the values of the matches with the factorial structure, and the high experience group had a lower

mean than the other two groups. An ANOVA will still be performed to further investigate the statistical significance of those differences.

Figure 6.

Boxplot, with low, medium and high chatbot experience groups as the independent variable and the percentage of matches with the original factorial structure as the dependant variable. The "Low" group includes 10, the "Medium" group 6, and the "High" group 7 participants.



Assumptions of the ANOVA were met, except for the normality distribution ($W = 0.9416$, $p > .05$). As the Assumption of Normality was not met, the nonparametric Kruskal-Wallis test was used to investigate the statistical significance of group differences in this sample. The result of the Kruskal-Wallis test suggests that there are no statistical-significant differences regarding matching the factorial structure between the three chatbot experience groups ($H(2) = 0.2806$, $p > .05$). This shows that previous chatbot experience does not play a role in predicting accordance with the factorial structure from the original study.

4. Discussion

4.1 Construct and Face validity in the BUS-11

In regard to answering the first research question and investigating whether the BUS-11 has sufficient construct and face validity, the results indicate that the scale has sufficient

construct and face validity. Especially regarding factors 1 (items 1 and 2), 4 (item 10) and 5 (item 11), the construct and face validity could be confirmed, as the results were in accordance with the original factorial structure. Factors 2 and 3 also displayed sufficient construct validity, as the items belonging to these factors were also sorted to said factors in the card sorting. Hereby, the items were differently distributed across the two factors compared to the original study. Additionally, item 7 was the only item belonging to factors 2 and 3 that fully matched the factorial structure in the card sorting results.

Based on the results, it can be said that the BUS-11 provides a good measurement for chatbot satisfaction. Especially the accessibility (factor 1), security/privacy (factor 4) and the time response (factor 5) of chatbots were perceived by the participants in accordance with the factorial structure. Regarding factors 2 and 3, participants did not agree as much. This still is in accordance with the original study and the factorial structure, as a strong correlation between factors 2 and 3 was found there. Even though factors 2 and 3 also show good construct validity, when considering their previous found correlation, there might be a possibility to combine both factors, which might give a more accurate representation of the mental model of people regarding chatbot satisfaction.

As previous research found a four-factor model to be more fitting (Waldmann, 2021), this might indicate that factors 2 and 3 can indeed be combined, as the other three factors displayed excellent construct and face validity and were frequently matched by the participants with the factorial structure. Additionally, the items in factors 1, 4 and 5 were often sorted matching the factorial structure, which is in accordance with the previous literature, as they also found the items belonging to the factors to have high factor loadings (Lopez & Borsci, 2021).

In conclusion, the scale provides a good estimate for chatbot satisfaction and therefore is also important in assessing chatbot usability. The hypothesis that the scale has sufficient construct and face validity can be accepted. The high correlation between factors 2 and 3 found in the original study and the four-factor model results from the other literature suggest a possible combination of factors 2 and 3.

4.2 Effect of Chatbot Experience and Card Sorting results

The second research question was: “Does previous chatbot experience affect the card sorting results, and to what extent these results match the expected factorial structure?”. The results indicate that previous chatbot experience has no significant effect on how the cards

were sorted. It seems that there is less agreement amongst participants with higher experience levels. This difference in distribution was mainly observed in the items belonging to factors 2 and 3. Additionally, item 4 was not grouped with other items in the medium experience level group, which contrasts the original study and the results from the overall sample. The good construct and face validity concluded from the first research question were also found in all three groups. Again, factors 1, 4 and 5 had very good construct validity, while the items belonging to factors 2 and 3 showed the same distribution level as in the overall sample.

The difference in distribution across the three groups (higher chatbot experience = less agreement on factorials structure) may indicate that the mental model of chatbot satisfaction becomes more individual with more experience and, therefore, knowledge on chatbots. As these differences were mainly observed regarding the items belonging to factors 2 and 3, they can also be attributed to the high correlation between these factors found in the original study. Regarding the isolation of item 4 in the medium experience group, it can be said that this might be due to the small sample size, as previous research and also the results from the overall sample suggest that item 4 frequently refers to the same factor as items 6 and 8 (in this study) and items 3 and 5 (in the original study). Therefore, the interpretation of further research regarding item 4 and the implication of such research on the factorial structure can be disregarded. Overall, no significant effect of chatbot experience on the card sorting results, and therefore on construct and face validity, could be observed.

This is in line with previous research, which found that experience had an effect on the usage of technology, but no effect on the evaluation (Shih et al., 2006). Similar to the overall sample, the distribution of agreement regarding items in factors 2 and 3, combined with the fit of the four-factor model resulting from previous studies (Waldmann & Borsci, 2021), further suggests the possible combination of factors 2 and 3. Further previous research on the effect of previous chatbot experience on the assessment of the BUS-11 does not exist.

In conclusion, chatbot experience seems to have no significant effect on the participants' mental model regarding chatbot satisfaction. Based on this, it can be said that the BUS-11 scale has good construct and face validity across different levels of chatbot experience. This is an important result, as this might have implications for the practical use of the scale. As the assessment of chatbot-user satisfaction is independent of chatbot experience, the scale can be used to assess a wider population, as experience has no influence. Especially

as the older population might be not as familiar with chatbots, the scale can be used to further develop chatbots according to their needs.

The lack of effects of people's experience on card sorting was also confirmed by a Kruskal–Wallis test. As most participants did match around 70% of the original factorial structure, it can be concluded that the factorial structure does not completely represent the mental model regarding chatbot satisfaction. This again indicates, combined with the results from the previous research questions, that factors 2 and 3 could be combined. This is also in accordance with the previous literature, which found a four-factor model more fitting (Lopez & Borsci, 2021).

4.4 Limitations

The closed card sorting design chosen in this study does not leave much room for the participants to sort the cards according to their mental model. This has limitations regarding the interpretation of factors 2 and 3, as these highly correlate, which was also observed in the card sorting results. This might be limiting, as there was no possibility to provide suggestions (e.g., for combining the two factors) for the participants, which may have brought valuable insights about their mental model regarding chatbot satisfaction. Additionally, the closed card sorting design limits the measurement of construct validity, as the participants cannot sort the cards according to their mental model but according to given constructs. Another possible limitation is the Dutch or German heritage of all participants. It is possible that a study in a different cultural context would yield different results. Additionally, it might yield different results regarding the effect of previous experience if a more complex study design is used, which is also more fitting to assess this topic, as the linear regression and the group analysis in this study gave limited insights on how experience influences chatbot usability. The low sample size also limited the ANOVA analysis's statistical power, which takes validity away from the results. Additionally, there was not much distribution regarding the age of participants, which may also limit the validity of the results, especially regarding the experience level, as the older population might still evaluate chatbot satisfaction differently, also compared to the participants with low experience.

4.5 Further research

Further research is needed to investigate whether cultural and age differences have an effect on the card sorting results. Especially age might provide valuable information for the practical use of the scale, as the older population might be less open to chatbot use, and

improvements in chatbot satisfaction and usability might make it more accessible. Additionally, it could be valuable to further investigate the relationship between factors 2 and 3 and research whether they might be combinable. This might give a more accurate representation of people's mental model regarding chatbot satisfaction. Further research regarding the ANOVA is also needed, as it was very limited due to the small sample size in this study. Further research might still find differences between groups with different levels of chatbot experience regarding grouping the cards in accordance to the factorial structure.

5. Conclusion

It can be concluded that the BUS-11 has good construct and face validity and that the factorial structure could be reproduced in this study. Additionally, chatbot experience did not have an effect on the card sorting results, indicating that there are no differences in the perception of chatbot experience regarding different levels of experience. Even though the factorial structure was confirmed in this study, it was indicated by the distributions of agreement levels regarding the items in factors 2 and 3 that these factors can be possibly combined. Overall, the BUS-11 provides a good measurement for chatbot satisfaction.

Reference List

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

Businesswire. (2019). The Global Chatbot Market is Forecast to Reach $5.63 Billion by 2023 – Asia Pacific to Witness the Highest Growth - ResearchAndMarkets.com. Businesswire. Retrieved 20th of August from https://www.businesswire.com/news/home/20190314005364/en/Global-Chatbot-Market-Forecast-Reach-5.63- Billion

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In Proceedings of the International

Conference on Information Systems (ICIS), 1-13. Retrieved from https://www.researchgate.net/profile/UlrichGnewuch/publication/320015931_Towards_Designing_Cooperative_and_Social_Conversational_Agents_for_Customer_Service/links/59c8d1220f7e9bd2c01a38a5/Towards-Designing-Cooperative-and-Social-Conversational-Agents-for-CustomerService.pdf

Io, H. N., & Lee, C. B. (2017). Chatbots and conversational agents: A bibliometric analysis. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. https://doi.org/10.1109/ieem.2017.8289883

ISO (2018). Ergonomics of human system interaction – Part 11: Usability: Definition and concepts (9241). Retrieved from https://www.iso.org/standard/63500.html

ISO (2019). Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems. Retrieved from https://www.iso.org/standard/77520.html

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

Lopez, S. M. K., & Borsci, S. (2021). Confirmatory Factor Analysis of a new Satisfaction Scale for conversational agents and the role of decision making style. [Bachelor Thesis]. University of Twente, Enschede, The Netherlands.

Mogaji, E., Balakrishnan, J., Nwoba, A. C., & Nguyen, N. P. (2021). Emerging-market consumers' interactions with banking chatbots. *Telematics and Informatics*, *65*, 101711. https://doi.org/10.1016/j.tele.2021.101711

Olsen-Landis, C. (2021, January 13). *Card sorting: a powerful, simple research method - IBM Design*. Medium. https://medium.com/design-ibm/card-sorting-a-powerful-simple-research-method-9d1566be9b62

Shih, P. C., Muñoz, D., & Sánchez, F. (2006). The effect of previous experience with information and communication technologies on performance in a Web-based learning program. *Computers in Human Behavior*, *22*(6), 962–970. https://doi.org/10.1016/j.chb.2004.03.016

Van den Bos, M., & Borsci, S. (2021). Testing a scale for perceived usability and user satisfaction in chatbots: Testing the BotScale. [Master Thesis]. University of Twente, Enschede, The Netherlands.

Waldman, A., & Borsci, S. (2021). User satisfaction and trust in chatbots: testing the Chatbot Usability Scale and the relationship of trust and satisfaction in the interaction with chatbots. [Bachelor Thesis]. University of Twente, Enschede, The Netherlands.

Weizenbaum, J. (1983). ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *26*(1), 23–28. https://doi.org/10.1145/357980.357991

ZEMČÍK, M. T. (2019). A Brief History of Chatbots. *DEStech Transactions on Computer Science and Engineering*, *aicae*. https://doi.org/10.12783/dtcse/aicae2019/31439

Appendix

Appendix A

Original Factorial structure of BUS-11

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

4 - Perceived privacy and security	1. I believe the chatbot informs me of any possible privacy issues.
5 - Time response	1. My waiting time for a response from the chatbot was short.

Appendix B

B1 General Information and Informed consent

Q3 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 can drop out at anytime. 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, 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:

Lukas Schwemin

E-Mail: l.schwemin@student.utwente.nl

Tel.: +49 1773390062

For any other questions or complaints, contact:

Jule Landwehr

E-Mail: j.landwehr@utwente.nl

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

Consent Q1 **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.

- ○ Yes (1)
- ○ No (2)

Consent Q2 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.

- ○ Yes (1)
- ○ No (2)

Consent Q3 I understand that taking part in the study involves me filling in a usability scale by myself and undergoing a closed card sorting test.

- ○ Yes (1)
- ○ No (2)

Consent Q4 **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.

- ○ Yes (1)
- ○ No (2)

Consent Q5 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.

- ○ Yes (1)
- ○ No (2)

Consent Q6 **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.

- ○ Yes (1)
- ○ No (2)

Q12 **Study contact details for further information**:

Lukas Schwemin – l.schwemin@student.utwente.nl

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

B2 Demographic questionnaire

Demographics Q1 How old are you?

_____

Demographics Q2 **What is your current gender identity?** (check all that apply)

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

**This question is optional and could be skipped

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

- ☐ Man (5)
- ☐ Woman (6)
- ☐ Female-to-Male (FTM/Transgender Male/Trans Man (7)
- ☐ Male-to-Female (MtF/Transgender Female/Transgender Woman (8)
- ☐ Genderqueer, neither exclusively male or female; (9)
- ☐ Additional Gender Category/(or Other), please specify (10) _____
- ☐ Decline to answer (11)

Demographics Q3 what is your nationality?

- ○ Dutch (1)
- ○ German (2)
- ○ Other: (3) _____

Demographics Q4 What is your level of English proficiency?

- ○ B1 - Intermediate (1)
- ○ B2 - Upper Intermediate (2)
- ○ C1 - Advanced (3)
- ○ C2 - Proficient (4)

B3 Previous Chatbot Experience

Q31 How familiar are you with chatbots and/or other conversational interfaces?

	Not familiar at all (1)	Slightly familiar (2)	Moderately familiar (3)	Very familiar (4)	Extremely familiar (5)
Indicate here: (1)	○	○	○	○	○

Experience Q2 Have you used a chatbot or a conversational interface before?

	Definitely not (1)	Probably not (2)	Possibly (3)	Probably yes (4)	Definitely yes (5)
Indicate here: (1)	○	○	○	○	○

Experience Q3 How often do you use chatbots weekly?

	0 times (1)	1-2 times (2)	3-4 times (3)	5-6 times (4)	Daily (5)
Indicate here: (1)	○	○	○	○	○

B4 Chatbots used in the survey

B4.1 information about chatbots

Q37 After this slide, you will be asked to fill in some questions regarding chatbots. Just for your information you will be given an short explanation of what chatbots actually are. Chatbots, or also known as conversational interfaces, are intelligent conversational applications that can mimic a

human conversation by engaging in voice and/or text output and input. They can serve many purposes such as in health care, customer service and also entertainment. An example of a popular chatbot is Siri. Hopefully this makes the concept of a chatbot more clear to you.

Good luck!

B4.2 First Chatbot

Q214 Chatbot: A&O Hostels

Go to the website where you can find the chatbot:

https://www.aohostels.com/en/

Q216

Perform the following task using the chatbot:

You want to go on a holiday to Berlin. You want to know how much a parking space costs in Berlin (Friedrichshain).

Q33 How much does a parking ticket cost in Berlin (Friedrichshain)?

_____

Q34 Were you able to complete the task?

- ○ Yes (1)
- ○ No (please specify why not) (2) _____
- ○ I am not sure (4)

B4.3 Second Chatbot

Q24 Chatbot: ATO

Go to the website where you can find the chatbot:

www.ato.gov.au

Q25

Perform the following task using the chatbot:

You moved to Australia from the Netherlands recently. You want to know when the deadline is to submit/lodge your tax return when doing it yourself.

Q39 When is the deadline for lodging your own tax return?

_____

Q41 Were you able to complete the task?

- ○ Yes (1)
- ○ No (please specify why not) (2) _____
- ○ I am not sure (3)

B5 Card sorting Instructions

Q27 **Card Sorting**

The last step is to do a card sorting test. This means that you must match the items of the scale that you have filled in into certain categories. These items are displayed on the left of your screen and the categories on the right. You can drag the items to the category you think the item belongs to. Please match the item to the category you feel like is most logical. Just to give you an example: when you have a category of 'clothing', you would match items such as 'dress' and 't-shirt' to this particular category. Good luck! (You can drag the cards to one group, scroll down and pick it up again, in case you have problems reaching the group with your cursor :))

Appendix C

R code used in Data Analysis

C1 Heatmaps

Needed Libraries

```
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_conflicts() ──

## ✕ 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/talgalili/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(dendextend))
## ----------------------
##
## Attaching package: 'dendextend'
## The following object is masked from 'package:stats':
##
##     cutree
library(pheatmap)
```

Read the data file

```
example_data <- read.csv("/Users/jule/Documents/Utwente Lecturer/Bachelor Thesis/Data Analysis/R_Datasheet.csv", comment.char="#")
rnames <- example_data[,1]
```

Transform data in numerical format and give names

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

Define colors of heatmap: red for high numbers

```
my_palette <- colorRampPalette(c("yellow","red"))(n = 299)
#Heatmap & Dendrogram
heatmap.2(dendrogram = "row", mat_data1, col = my_palette, density.info="none", trace="none",
          revC = TRUE, main="", cexCol = 1, cexRow = 1, margins = c(5, 5))
```

C2 Effect of Chatbot Experience on Accordance with Factorial Structure

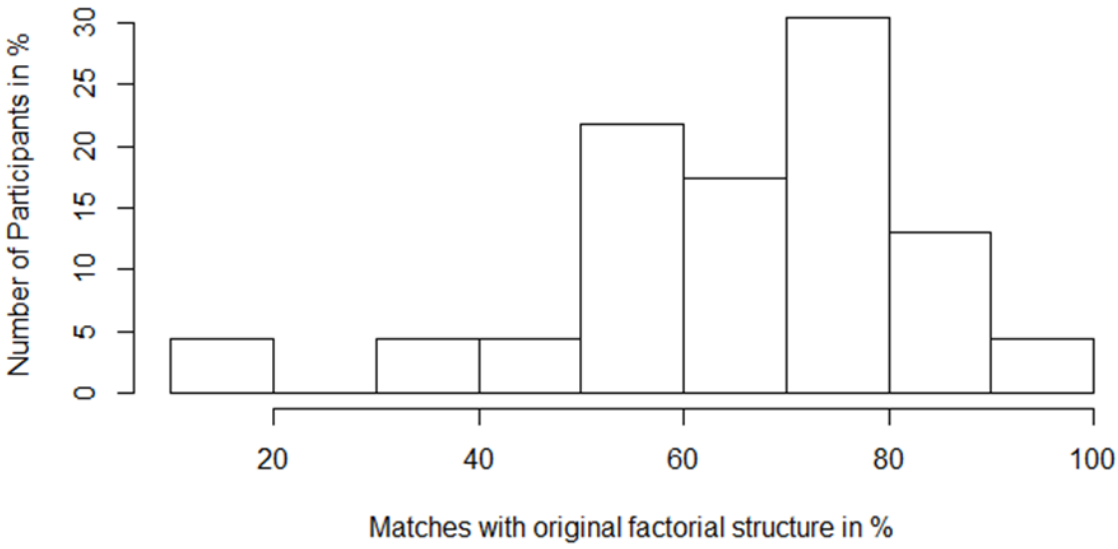
```
#Read the file
```

```
crop.data <- read.csv("C:/Users/HP/Documents/Module 11/Bachelor/Data analysis exel/Final total scores.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "factor", "numeric"))

#Histogram

> h = hist(crop.data$Total_per, plot = FALSE)
> h$counts = h$counts/sum(h$counts)*100
> plot(h, xlab = "Matches with original factorial structure in %", ylab = "Number of Participants in %", main = "")

#Recoding Group variable

 > crop.data$Groups <- dplyr::recode(crop.data$Groups, '1' = "Low", '2' = "Medium", '3' = "High")

#Boxplot

boxplot(crop.data$Total_per ~ crop.data$Groups, xlab = "Level of Chatbot Experience", ylab = "Matches with original factorial structure in %", main = "")

#Shapiro-wilk test
shapiro.test(crop.data$Total_per)

#Lavene test
leveneTest(Total_per ~ Groups, data = crop.data)
```

Appendix D

D1 Assumption of Normality

Histogram to test normal distribution in the sample regarding matches with the factorial structure in %.



Shapiro-Wilk test of normality

W = 0.94159, p-value = 0.1943

D2 Assumption of Homogeneity of Variance

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	1.0396	0.3719
	20		