23-01-2023

Conversational Agents: Using Open Card Sorting to Evaluate the Factor Structure of the BUS-11

Bachelor Thesis

Chiara Emma Wermter
S2364956

First supervisor: MSc Jule Landwehr
Second supervisor: Dr. Simone Borsci

University of Twente

Contents

Abstract

This paper examines the factor structure of the Bot Usability Scale (BUS) as proposed by Borsci et al. (2022). The BUS aims to assess the user satisfaction with chatbots (Borsci et al., 2021), which are systems that mimic natural language to interact with users (Rese et al., 2020). Thereby the BUS incorporates aspects of communication, as well as the user's perception of data security, making it more suitable for the assessment of chatbots than commonly used usability scales (Borsci et al., 2021). Recent research has indicated that the recent version of the BUS-11, has only four instead of five underlying factors (Borsci & Schmettow, 2023). This study investigates the factor structure, with a new approach by deriving factor models from the mental models of users. For the identification of mental models an open card sorting was conducted. The results included the mental models of 47 participants and were assessed with heatmap analyses. The newly created models and the Four-Factor Model by Borsci & Schmettow (2023) were additionally, evaluated with a confirmatory factor analysis. The results indicated that the Four-Factor Model proposed by Borsci & Schmettow (2023) could be identified during the heatmap analysis if a lenient threshold of 30% agreement was applied. The confirmatory factor analysis suggested that the Four-Factor Model by Borsci & Schmettow (2023) was not fitting the data used for the confirmatory factor analysis. Two of the three novel models showed an acceptable fit to the data. Based on the Akaike information criterion and Bayesian information criterion, a hierarchical model with four first-order factors and thee second-order factors provided the best fit for the data.

Conversational Agents: Using Open Card Sorting to Evaluate the Factor Structure of the BUS-11

In recent years, technology and the internet developed with increasing speed. To make products accessible and user-friendly companies increasingly make use of natural language, i.e., language that has naturally evolved in humans (e.g., speech, text) (Rese et al., 2020). Software systems that engage with users by mimicking human interaction, thus using natural language, are called conversational agents (Radziwill & Benton, 2017). Popular examples of conversational agents (CA) include systems like Siri and Alexa. Chatbots, another type of CA, are frequently integrated into websites. They belong to disembodied CAs, as they usually do not have an animated body or face. Therefore, they primarily engage with individuals using text messaging interfaces (Araujo, 2018). While chatbots are sometimes employed for entertainment purposes (e.g., chit-chat chatbots), they are more commonly used to support users in completing certain tasks (Li et al. 2016).

Task-oriented chatbots are employed to fulfil a goal. Thereby, chatbots can provide customers with fast and accessible information and help. They lower the threshold for inquiries, as chatbots are perceived as non-judgemental and do not elicit time pressure in the customers (Følstad et al., 2018). Companies employ chatbots to reduce the number of user requests requiring human attention and to reduce the response times for inquiries (Xu et al., 2017). Overall, chatbots allow users to solve problems or gather information fast and independently without occupying a company's human resources**.**

In order to reach their potential, chatbots and the surrounding web interface must be well designed. The webservice's usability of a web service has a positive influence on customer satisfaction and thereby increases the purchase intention (Bali et al., 2008). Regarding chatbots, Følstad & Brandtzaeg (2020) found that especially pragmatic aspects of user experience are essential for a successful implementation. Interactions with chatbots should be useful and efficient. Thereby, it is essential for the chatbot to understand the user's intentions correctly (Følstad & Brandtzaeg, 2020; Følstad & Taylor, 2021). Accordingly, usability is a determining factor for the successful implementation of chatbots.

Usability is "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO, 2018). Although many established questionnaires are available for usability assessment (e.g., CSUQ, SUS and UMUX) (Lewis, 2018), established usability questionnaires might not be suitable for evaluating conversational agents. Borsci et al. (2021)

point out that in contrast to other interactive systems, the functions of chatbots heavily rely on conversational aspects. Commonly used usability scales were not developed to assess systems that rely on natural language and might not be equipped to evaluate chatbots.

This issue becomes apparent when reviewing frequently reported issues with chatbots, as those often concern the quality of conversation. Among the documented issues are vague responses (Li et al. as cited by Li et al., 2016), a display of inconsistent personality (Li et al., 2016) as well as brief, unsatisfying answers or answers that disregard priorly given information (Vinyals & Le, 2015). Many chatbots are not equipped to consistently handle user requests as they lack understanding of the user's intention. Due to the absence of items about conversation quality, established usability scales would fail to identify the abovementioned issues without additional information.

To create a standardised tool that assesses user satisfaction while incorporating the evaluation of conversational aspects, Borsci et al. (2021) created the BOT Usability Scale (BUS). Satisfaction can be described as a positive attitude towards and comfort with using the system (ISO as cited in Frøkjær et al., 2000). The BUS scale was created by evaluating and narrowing down attributes found in the literature and adding relevant attributes identified in a study, which resulted in the BUS-15 (Borsci et al., 2021). Analysis of the BUS-15 indicated that the scale had five underlying factors (Borsci et al. 2022). The BUS-11 is the most recent version of the BUS and comprises 11 items (Borsci & Schmettow, 2023). The most recent version of the BUS-11 with a Four-Factor structure can be seen in Table 1. The Four-Factor Structure was identified by Borsci and Schmettow (2021) by applying a design-o-metric approach that was proposed by Schmettow (2021).

Table 1

Table showing the Four-Factor structure of the BUS-11 with a description of the items

Factor	Item
1 Accessibility - Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable.
	2. It was easy to find the chatbot.
2 Functional interactive conversation-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.
	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.
3 Privacy - Perceived privacy and security	10. I believe the chatbot informs me of any possible privacy issues.
4 Responsiveness - Time response	11. My waiting time for a response from the chatbot was short.

The design-o-metric approach addresses the issue that analysing designs using traditional psychometric approaches is not focusing on the characteristics of the designs but rather on the participants. Therefore, a design-o-metric analysis was conducted by averaging over participants and performing a confirmatory factor analysis. Priorly, the factor structure of the BUS was evaluated by averaging over the designs of the chatbots, which resulted in a Five-Factor Structure (Borsci et al., 2021; Borsci et al., 2022). The reassessment of the BUS-11 showed that the factors "perceived quality of chatbot function" and "perceived quality of conversation and information provided" could be merged into one factor. The new factor was named "Functional interactive conversation-Perceived quality of chatbot function" and is constituted of items three to nine. Until now, the factor structure of the BUS-11 was mainly analysed using different approaches to factor analysis. The factor analysis of the Five-Factor Model and the Four-Factor Model showed that both models sufficiently fit the data (Borsci & Schmettow, 2023).

Therefore, exploring the factor structure using a different method could provide valuable insight. Beerlage-de Jong et al. (2020) describes an approach in which a factor structure is identified by using mental models. Thus, the mental model of individuals can be used to deduct a model and test its fit to BUS-11 data. Mental models are a person's knowledge structure (Schmettow & Sommer, 2016) or "a person's simplified representation of how something works in the real world" (Beerlage-de Jong et al., 2020). Individuals have mental models about their surrounding world and everything in it. They can represent a simplified framework or explanation of how something works or is structured. Hence, individuals could structure the BUS-11 items into groups of their liking resulting in a structure that can be used to build a factor model. This empirical method is called open card sorting and can be used to collect data about the mental model of individuals (Schmettow & Sommer, 2016). In this case, an open card sort should be preferred to a closed one. In an open card sort, participants can create and name categories that correspond to their original mental model (Beerlage-de Jong et al., 2020). In a closed card sort, the participants must sort the items into predefined categories, which may not correspond to their natural model.

Identifying how individuals structure the items of the BUS-11 might provide more insights into their approach to user satisfaction. Of particular interest is whether the mental model is consistent with statistically found models and whether the mental model has more in common with the psychometrically identified model or the design-o-metric model.

In this study, participants’ mental models are explored using an open card sort. The goals of this research are to (1) compare the mental model of the participants to the Four-Factor Model (Borsci & Schmettow, 2023) and assess the face validity (2) find the underlying mental model participants use for the BUS-11, and (3) assess the construct validity and fit of the found models and the Four-Factor Model (Borsci & Schmettow, 2023) using confirmatory factor analysis.

Methods

Participants

In total, 58 participants took part in the study. All participants that did not fulfil the language requirements of having at least basic language proficiency, only partook in the first part of the study or did not follow the instructions were excluded. Six participants misinterpreted the card sorting instructions (2 English, 3 German, 1 Dutch) and used the groups to rate the previously evaluated chatbot. Further, it appeared that one participant did the card sorting three times, as the categories had identical names and items. For the analysis, two of the identical results were removed.

In total, 47 observations were used in the analysis (26 English, 14 German, and 7 Dutch). Approximately 66 % of the participants were female (Age mean: 22.11; SD: 6.38). Participants were recruited using the University of Twente's SONA system and social media. Ethical approval was obtained from the University of Twente's Ethics Committee before the participants' recruitment started. Furthermore, two pilot studies have been conducted to identify potential issues before the study's publication. A total of three German participants took part in the first pilot (two male; Age mean: 24.33; SD: 4.73). All of them had at least basic language proficiency. In the second pilot study, two German participants participated (one male, Age mean: 20.5, SD: 2.12). Both participants had sufficient language skills.

Materials

Platforms and languages

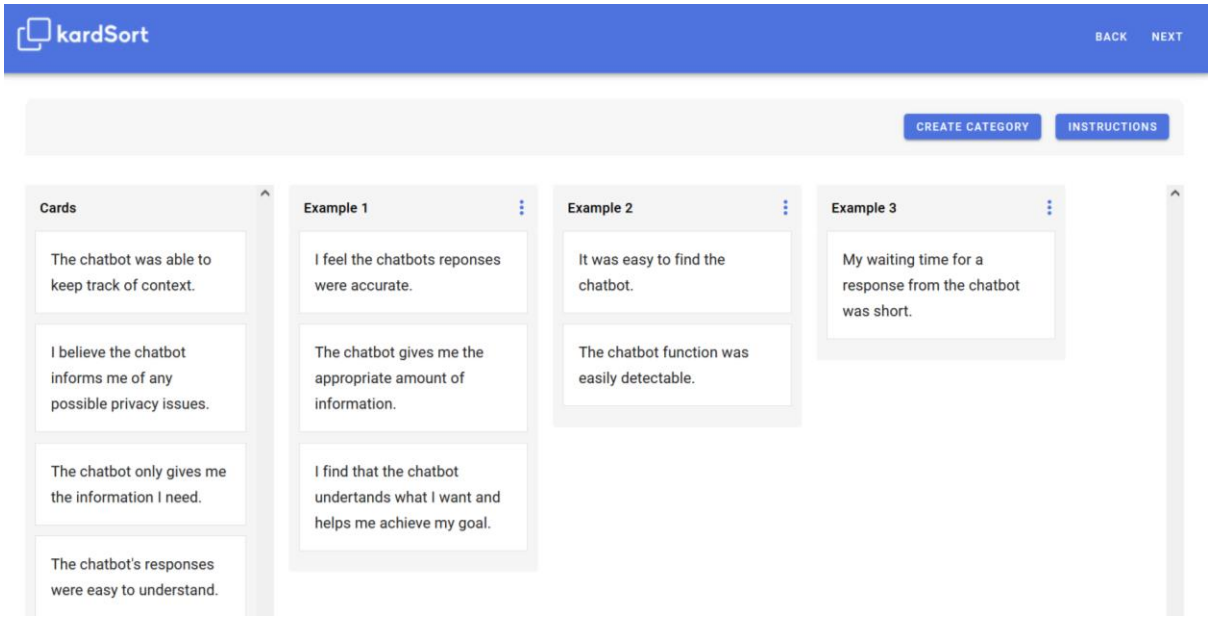
The study was split up between two platforms. The demographics, the chatbot task and evaluation with the BUS-11 were conducted on Qualtrics. Since Qualtrics is not equipped to conduct open card sorts, the participants were instructed to visit kardSort.com to partake in the second part of the study. A depiction of the kardSort interface can be found in Figure 1. The cards to be sorted can be found on the left of the interface, the created categories can be found on the right side next to it. To connect the Qualtrics responses to the kardSort responses, Qualtrics generated 7-digit IDs that participants were required to enter before the card sorting. The first five digits of the ID consist of a random number between 10.000 and 99.999 that Qualtrics generates. The sixth and seventh digits are the age of the participant.

Further, both study parts were available in English, Dutch, and German. In Qualtrics, the survey language was automatically set to the language used in the browser and could

alternatively be changed in a drop-down menu on the top right of the survey. For the second part of the study, participants were directed to the kardSort study in the respective language.

Figure 1

Picture of the Card Sorting Interface



Qualtrics.

Informed consent

An informed consent form is provided to participants before the start of the study. The participants are informed about data processing, purpose, content of the research, use and future use of information. In addition, the participants were provided with contact information.

Demographics

Inquiries about demographics are made in the first part of the survey. The demographic part of the survey includes questions about age, gender identity, sex, nationality, language proficiency and prior experience with chatbots. Further, the participants were asked whether they are diagnosed with attention deficit hyperactivity disorder (ADHD) or attention deficit disorder (ADD) and whether they receive medication for those disorders. The entire Qualtrics survey can be found in Appendix B.

Chatbot example and task

To familiarise participants with chatbots and the BUS-11, they were given a task to solve using a chatbot. Participants were informed that they could continue the study even if they could not solve the task and did not have to provide personal information to the chatbot (see Appendix B). Overall, this part included a question about the task, whether participants were able to complete the task and the assessment of the example chatbot with the BUS-11. Participants were given the link to the chatbot (www.oxxio.nl/klantenservice) and a brief explanation of how to change the chatbot's language (German and English versions) to solve the task. Participants were asked to fill in the BUS-11 items on a 5-point Likert scale (strongly disagree – strongly agree). The items were shown in random order, allowing participants to familiarise themselves with the scale without being biased by the order of the items.

kardSort

Procedure

The participants were informed that they must be at least 18 years old and speak English, German, or Dutch to participate in the study. Further, participants were strongly advised to use a PC or Laptop for their participation, as the first pilot study suggested that the mobile version of kardSort.com has poor usability. After reading the instructions, the participants were informed about informed consent. If participants gave consent, they were asked to fill in the survey about their demographics and their experiences with chatbots. In the next step, participants were asked to visit the page https://www.oxxio.nl/klantenservice and to inquire about the advantages of a Smart meter. To answer the question, participants could pick four multiple-choice options, of which two were right. Afterwards, the participants were asked whether it was possible for them to finish the task and to fill in a BUS-11 with a randomized item order. After the evaluation with the BUS-11, the participants were given their ID and redirected to the kardSort website. Participants had to enter their ID receiving the card sorting instructions. For the card sorting participants created groups with items, which they named. After the card sorting, the study ended, and the participants were thanked for their participation.

Data analysis

Excluding participants

The data was exported from Qualtrics and kardSort. SynCaps Version 3 and R Studio Version 2022.07.2 were used for the analysis. To ensure clear documentation, R-Markdown was used.

One of the main goals during this step was to create a script to prepare data more efficiently. Without automatising the process, four datasets would have to be viewed individually to match the Qualtrics response with the corresponding kardSort response. In addition, a separate ID that kardSort (kardSort ID) used in the SynCaps file had to be identified to prepare data for SynCaps. With the R script documented below, the kardSort ID, Qualtrics ID and language of participants can be viewed in one data frame. Thus, participants that did not complete the card sorting could be identified and removed from the Qualtrics dataset more easily. Further, all IDs of participants that did not meet certain criteria (e.g., language requirement) could be displayed, allowing to identify participants based on their Qualtrics responses in the SynCaps datafile.

The necessary data was exported from Qualtrics and kardSort. To retain the variable and value labels the Qualtrics dataset is exported as sav-file and transformed after a codebook was generated. The Qualtrics translations feature allowed exporting data from all participants in one dataset regardless of the language chosen for participation. During the preparation for analysis, unnecessary data was removed, a codebook was created, and variables were recoded.

In the kardSort application, one survey had to be created for each language. For each survey on kardSort, several datasets could be exported. The dataset compatible with SynCaps contained the cards and created groups as well as a participant number generated by kardSort (kardSort ID). The kardSort ID numbered the participant from the latest to the earliest. The ID generated by Qualtrics (Qualtrics ID), that was used to connect the participant to their survey response, could only be found in a kardSort dataset called “Questionnaire responses”. To identify and exclude participants, the Qualtrics ID had to be matched with the kardSort ID derived from the three kardSort surveys. Therefore, part of preparing the data was to match the Qualtrics ID with the corresponding card sorting results.

To automatize this progress, a script in R was used (see Appendix C). The functionality of the script is depicted in Appendix A. The script loaded the three datasets from kardSort that contained all Qualtrics IDs. The kardSort ID was generated for each kardSort dataset by numbering the participants. Afterwards, the three kardSort data frames

were merged into a new data frame that contained both IDs for all participants. The resulting data frame was merged with the Qualtrics data by the Qualtrics ID. The data from participants whose Qualtrics ID was not present in both data frames was moved into a separate data frame. Moreover, the script allowed matching the Qualtrics responses of participants with their data in SynCaps.

Heatmap analysis of the Card Sorting data

Words association data of the card sorting are analysed using the Jaccard coefficient, which assesses the similarity of items. The Jaccard coefficient is generated by dividing the number of groups in which both items were placed by the number of in which either item was grouped (Schmettow & Sommer, 2016). The Jaccard scores of all participants were then combined in an Item-to-Item matrix. Participants who did not correctly interpret the card sorting instructions were excluded from the final Item-to-Item matrix. The heatmaps were generated in R using the package ComplexHeatmap.

The heatmaps are used to compare the combined mental models of participants with the Four-Factor Model proposed by Borsci & Schmettow (2023) and to explore new models. To facilitate the comparison between the Four-Factor model (Borsci & Schmettow, 2023) and the participants' metal model, the first heatmap was plotted unclustered. In the resulting heatmap, the items are ordered according to the BUS-11 order. In addition, it is determined which threshold must be applied so that the Four-Factor Model by Borsci & Schmettow (2023) can be extracted. For the extraction of own factor models, two additional heatmaps were plotted. To facilitate the extraction of models, the rows, and columns of the heatmap were clustered. The clustering causes rows and columns of frequently grouped items to be displayed adjacent in the heatmap.

In order to extract several models, two different thresholds of the item-item-agreements were used. The stricter of the two stipulates that factors only consist of items that were grouped at least 50% of the time. Further, a more lenient threshold of a grouping frequency of at least 40 % was applied. Further, two types of points of interest are determined. All cells within the models whose values are below the threshold are outlined in light blue. The second type of points of interest includes all cells outside the factors that exceed the threshold.

Further, the data from the card sort is used to analyse the face validity.
Face validity describes "the degree to which a measure appears to be related to a specific construct in the judgement of non-experts" (Taherdoost, 2016). The face validity can be

evaluated by comparing the names of the factors with the names that participants gave identical groups. If the participants' category names resemble the factor names the scale has face validity.

Construct validity - Confirmatory factor analysis

To evaluate the fit of the factor models that have been deducted from the mental models and assess the construct validity of the Four-Factor Model under exploration by Borsci and Schmettow (2023), a confirmatory factor analysis was performed.

First, Cronbach **α** was calculated to assess whether the items of the BUS-11 measure the same underlying construct. Thereby, $\alpha \geq 0.7$ were considered acceptable (Urdan, 2017). The assumption of normality is assessed with the Shapiro-Wilk test and by calculating the skew and kurtosis of the BUS-11 total scores. A significant result of the Shapiro-Wilk test indicates that the data are not normally distributed (Ghasemi & Zahediasl, 2012). Further, the indications from Hatem et al. (2022) were used, to interpret the skew and kurtosis. A skewness of -0.5 to o.5 shows that the data is distributed fairly symmetrically. A skewness of 0.5 to 1 or -1 to -0.5 indicates that the data are moderately skewed. Lastly, a skewness of more than 1 or less than -1 indicates that the data are skewed highly. For kurtosis, a value close to 0 indicates a mesokurtic distribution. A kurtosis > 0 indicates a leptokurtic distribution, and with a kurtosis < 0 the distribution is platykurtic (Hatem et al. 2022). Therefore, a skewness between -0.5 and 0.5 and a kurtosis close to 0 would be indicating a normal distribution.

During the assessment of normality, it is important to consider that only a few chatbots have been evaluated with the BUS-11. It can be assumed that the scores of the different evaluated chatbots do not share the same distribution parameters, which could result in a non-normal distribution. In this case, the distribution might show several peaks as the distribution includes scores from chatbots with potentially varying quality.

Further, the BUS-11 uses a five-point Likert scale. Thus, the robust maximum likelihood method will be used for the CFA as it is equipped to analyse ordinal data (Li, 2016). First, to compare the model fit parameters to prior studies a confirmatory analysis was replicated using the same criteria as Borsci et al. (2022). For models to be accepted, the factor loadings of each model had to be above 0.6. Further, the Chi-square to the degree of freedom ratio should be below 3, the comparative fit index should yield results of 0.9 or higher, and the root mean squared error approximation (RMSEA) should be below 0.7. The standardized root mean square (SRMR) residual aims to be below 0.8.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are calculated to pick the most suitable model. For the evaluation, several models' AIC and BIC values are compared. The models with the lowest AIC and BIC values are the comparably better-fitting models. The AIC is recommended for selecting predictive models, and the BIC for selecting descriptive models. Nevertheless, the AIC tends to favour more complex models than the BIC (Cavanaugh & Neath, 2018). Further, the AIC may show negative bias in smaller samples (Cavanaugh, 1997, as cited in Cavanaugh & Neath, 2018). To account for the law of parsimony, models with both a comparably low AIC and BIC are preferred in the final selection.

Results

Heatmap

Three heatmaps plotted from the combined Item to Item matrixes are presented in the following sections. The first heatmap (see Figure 2) was used to analyse whether the clusters of the heatmap resemble the Four-Factor Model proposed by Borsci & Schmettow (2023). Hence, the items are sorted according to the BUS-11 scale. To facilitate the analysis, the cells that belong to factors according to the Four-Factor model are outlined in black in the heatmap.

In the second heatmap, a threshold of at least 50 % agreement was used to derive factor models from participants' mental models (see Figure 3). With the application of the 50% threshold, a Six-Factor model was found. The model is outlined in black.

For the third heatmap, a less restrictive threshold of at least 40 % agreement was used (see Figure 4). A found Four-Factor model is outlined in black.

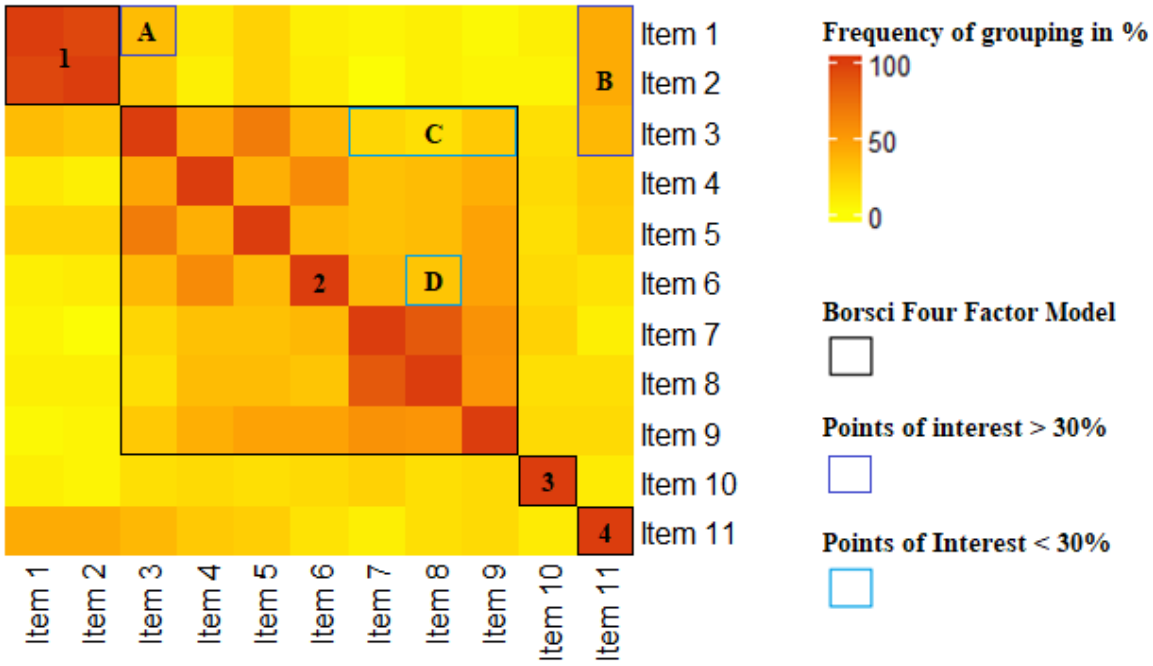
Each cell of the heatmaps represents the frequency in % with which two respective BUS-11 items were grouped. The colour scale of the heatmap ranges from pale yellow (low frequency) to deep red (high frequency of groupings). The heatmaps were plotted from the data of 47 participants. The items that have been sorted the most appear as clusters among the diagonal. In the figures, the factors are outlined in black and numbered.

To assess the face validity of the BUS-11, the group names used by participants who grouped the items identically to the Four-Factor Model were compared to the factor names proposed by Borsci & Schmettow (2023). The names of groups created by participants with identical items were reviewed for newly created factors. The names were picked to be specific enough to distinguish the concepts behind the factors from each other. To avoid

confusion, groups that were identical to the ones of the Four-Factor model by Borsci & Schmettow (2023) were named after the BUS-11 factors.

Figure 2

Unclustered Heatmap of the Card Sorting



Note: Unclustered heatmap of the combined Jaccard score of the open card sort. The numbered clusters, outlined in black show the Four-Factor Model by Borsci & Schmettow (2023). Two types of points of interest are outlined and labelled with letters. All cells within the models whose values are below the threshold of at least 30% agreement are outlined in light blue. Cells outside the factors that exceed the threshold of 30% are outlined in dark blue.

Un-clustered heatmap- Comparison to Four-Factor Model by Borsci & Schmettow (2023)

Overall, the Four-Factor structure as proposed by Borsci & Schmettow (2023) is visible in the heatmap. The analysis of the item-item matrix showed, that the Four-Factor model can only be deducted if a rather low threshold of at least 30% agreement is implemented. The factors of the Four-Factor model are outlined in black. Further, two different types of points of interest are outlined in blue. Cells within the model that show an agreement frequency below the threshold are outlined in light blue. Points outside of the model, that were sorted together by more than 30% of the participants are outlined in dark blue.

The factor Accessibility is clearly visible in the top left of the heatmap and is marked

with the number 1. Items 1 and 2 were sorted together by 45 out of 47 participants (95.74 %). Nevertheless, the points of interest A and B show that items 1 and 2 were also sorted together with the items 3 and 11. Item 1 was grouped together with item 3 in 34% of the cases. Item 11 was grouped with item 1 and item 2 by 42.55% of the participants. Fifteen participants formed groups identical to the Accessibility factor and either named it “Accessibility” or gave it names related to the concepts of Findability or Location. This indicates that the items measure the constructs, that they are supposed to measure. The factor Functional interactive conversation, marked with number two, contains items 3, 4, 5, 6, 7, 8 and 9. Even though the factor is visible in the heatmap, the cells of the points of interest marked as C and D show that not all of the items were grouped with each other with a frequency of at minimum 30%. At point of interest C, it appears that item 3 and item 8 were grouped 17.02% of the time, item 3 and item 7 were grouped 21.28% of the time and item 3 and item 9 were grouped 27.66% of the time. At point of interest D, the items 6 and 8 were grouped 29.79% of the time. In the heatmap, it appears that Factor 2 (“Functional interactive conversation”, see Figure 2) could be one broad overarching factor that describes several smaller factors. Items 3 and 5, 4 and 6 and 7,8 and 9 were frequently sorted into their own groups. Item 9, however, was regularly sorted with all Items of Factor 2, which is in accordance with the Four-Factor Model.

The items of the Factors Functional interactive conversation have only two participants grouped together. These named the factors "quality of answers in chatbot" and "information provided". Both names seem to be related to the concept of the factor. However, due to the small number of groups identical to the factor, no solid conclusions on content validity can be drawn.

Factor 3 (“Privacy”, see Figure 2) is clearly discernible in the heatmap and marked with “3” in Figure 2. Seventeen participants grouped item 10 alone. The names of the groups are all related to privacy, transparency, or concerns. This indicates that item 10 measures the construct that it is intended to measure.

Similarly, factor 4 (“Responsiveness”) composed of Item 11 and marked with “4” can be seen in the heatmap. The point of interest “B” indicates that Item 11 was frequently sorted together with the Items 1, 2 and 3. Item 11 was grouped with item 1 and item 2 42.55 % of the time. Item 3 and item 11 were grouped 36.17% of the time. Item 11 was sorted into its own group by nine participants. Seven of the groups were named time and waiting time. The remaining two were called "performance" and "advantage". It therefore appears as if the respective participants were able to derive the concept behind the item.

Clustered and ordered heatmap

For the extraction of own factor models two clustered heatmaps were plotted. Due to the clustering the rows and columns of items that were frequently grouped together are displayed adjacent to each other in the heatmap.

In the first heatmap (see Figure 3) a threshold of at least 50% agreement was used for the extraction of a factor model. The result was a Six-Factor model, which is outlined in black in the heatmap. Further, no points of interest were found with the application of the threshold.

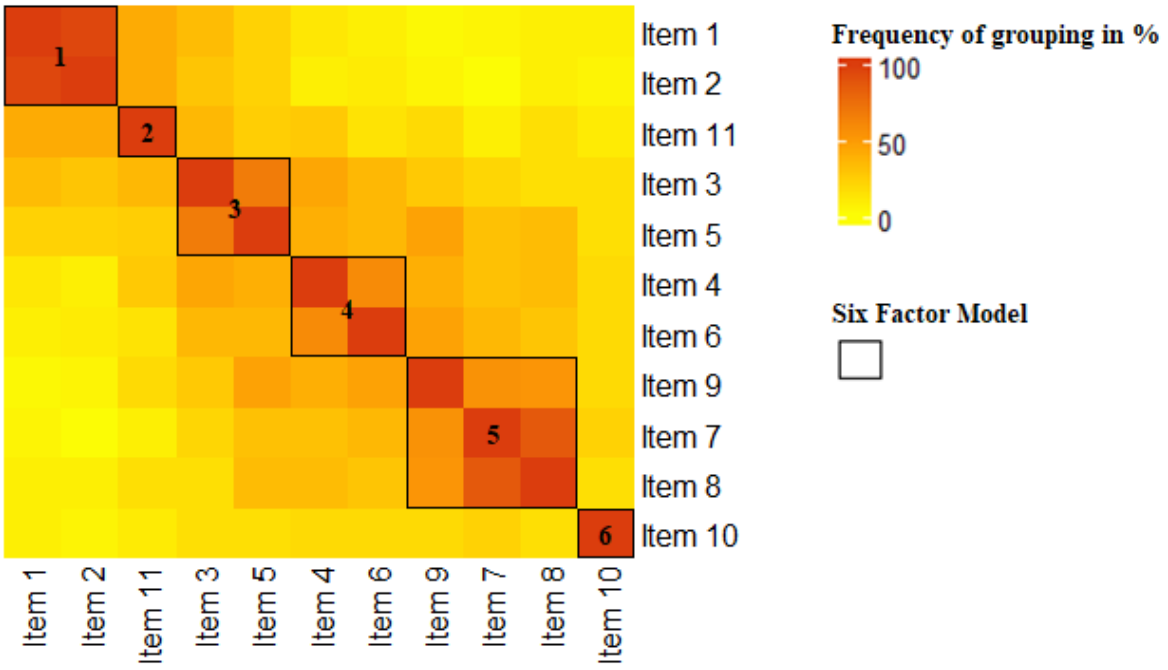
Further, a hierarchical factor model was created based on the Six-Factor Model and the Four-Factor Model by Borsci & Schmettow (2023). The four factors of Borsci's Four-Factor Model represent the first order factors of the model. The factor Functional interactive conversation is described by three second order factors that stem from the Six-Factor model.

Secondly, a less restrictive threshold of at least 40% agreement was used for the extraction of another factor model. Thereby, a Four-Factor model was found that is outlined in black in the second heatmap (see Figure 4). Further, two different types of points of interest are outlined in blue. The cells within the model with an agreement frequency below the threshold are outlined in light blue. Points outside of the model, that were sorted together in more than 40% of the cases are outlined in dark blue.

The names of the factors were picked by reviewing the names that participants gave groups that contain identical items. The aim was to select names that are specific enough to distinguish the concepts behind the factors from each other. Factors that contained the same items as Factors of the Four-Factor model were named after the model by Borsci & Schmettow (2023) to avoid having identical factors with different names.

Figure 3

Clustered Heatmap of the Card Sorting



Note: Clustered heatmap of the combined Jaccard score of the open card sort. The numbered clusters, outlined in black show the Six-Factor Model that was created during the analysis. The threshold for the model extraction was a minimum of 50% agreement.

Six-factor model

Accessibility- Items 1& 2

The first clearly defined cluster labelled with "1" is composed of Items 1: "The chatbot function was easily detectable" & Item 2: "It was easy to find the chatbot". Both items were grouped 45 times. The described cluster contains the same items as the factor "Perceived accessibility to chatbot functions" proposed by Borsci & Schmettow (2023).

Fifteen participants grouped Items 1 & 2 into one separate group. While participants used different names for the group, most names contained inflected forms of the word-finding (e.g. "Ease of finding", "Findability", "Finding chatbot function"). Further, the term "Accessibility" was used by three participants. In the following, the cluster will be called "Accessibility".

Quality of Chatbot expressions- Items 3 & 5

The cluster labelled "3" contains item 3 "Communicating with the chatbot was clear", and Item 5: "The chatbot's responses were easy to understand". Both were sorted together 32 times. The four participants who created groups with only items 3 and 5 called the resulting group "Comprehensibility of answers", "Quality of Chatbot expressions", and "Communication" (2). In Borsci's Four-Factor model, Items 3 and 5 belong to the factor "Functional interactive conversations" with the Items 4, 6, 7, 8 and 9.

Contextual understanding- Items 4 & 6

Another cluster contains Item 4: "The chatbot was able to keep track of context." and Item 6: "I find that the chatbot understands what I want and helps me achieve my goal.". The cluster is labelled with the number 4 in the heatmap. The items have been grouped 28 times (59.57 %) and seem to be concerned with the contextual understanding of the chatbot. Three participants sorted Cards 4 and 6 together and used the names "Chatbot understand context", "Usability" and "Contextual understanding" to describe the group. In Borsci's Four-Factor Model the Items 4 and 6 belong to the factor "Functional interactive conversations." together with Items 3, 5, 7, 8 and 9.

Quality of information in answer- Items 7, 8 & 9

The next cluster "5", contains the items 7, 8 and 9. Cards 7 and 8 have been grouped 41 times (87.23 %). Card 7 contains the item: "The chatbot gives me the appropriate amount of information". Item 8 is "The chatbot only contains the information that I need." Thus, both items are concerned with the amount of information, and their fit to the user's need.
Five participants created a group with only items 7 and 8 called the group: "Amount", "Content", "Amount of Information", "Information provided" and "Information". Card 9 was combined with Card 7 26 times (55.31 %) and with card 8 25 times (53.19 %). It contains the item: "I feel like the chatbot responses were accurate". Therefore, all items of this cluster are concerned with the chatbot's answer. Six participants created a group with items 7, 8 and 9. The groups were called "content", "Quality of information in answer", "given information", "Simplicity and accuracy", "Information" (2).

In Borsci's Four-Factor Model, items 7, 8 and 9 are in the same factor. However, Borsci's "Functional interactive conversations" factor also contains the Items 3, 4, 5, and 6.

Privacy- Item 10

The cluster labelled “6” on the bottom right of the heatmap (see Figure 3) only contains Item 10: “I believe the chatbot informs me of any possible privacy issues”. Only few participants sorted Card 10 together with other cards. Eighteen participants created a group with only item 10; of those, most named the group privacy.

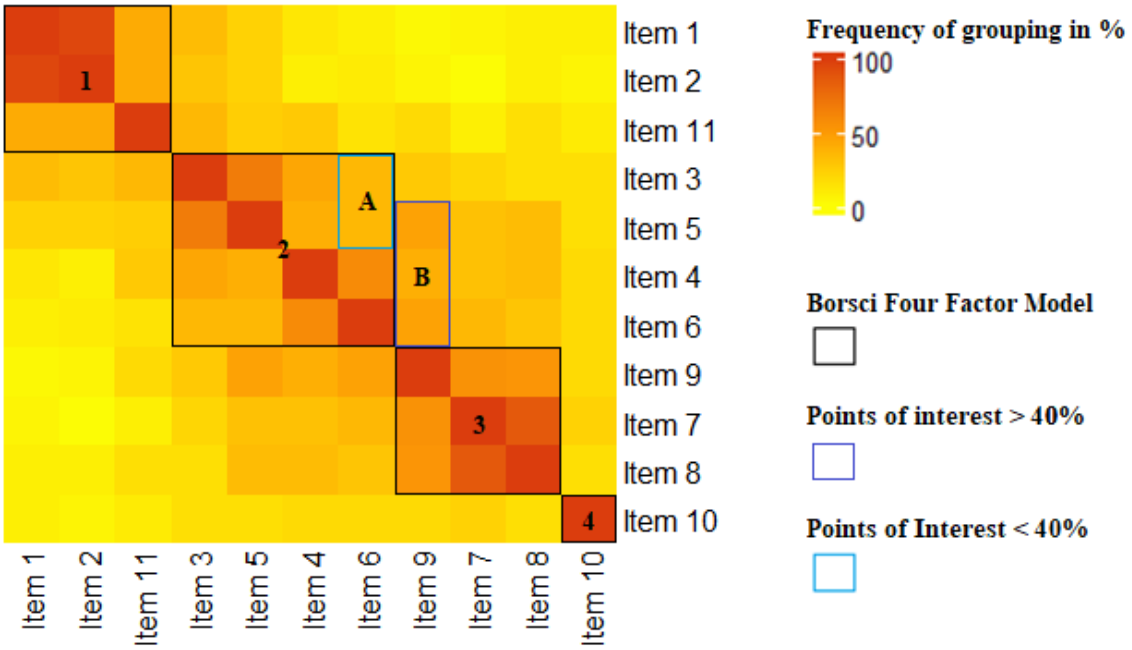
The finding that item 10 was mostly grouped on its own is in line with the Four-Factor Model proposed by Borsci & Schmettow (2023). Participants who did not group Item 10 in accordance with Borsci’s Four-Factor Model, mostly grouped it with the Item 4. Item 10 was grouped nine times with each Item 4, Item 6, and Item 9 (19.15 %). Both items 4 and 6 concern the chatbot’s ability to understand the users. Item 9 evaluates the accuracy of the chatbot’s responses.

Waiting time-Item 11

Lastly, Item 11: “My waiting time for a response from the chatbot was short” constitutes its own factor, labelled as “2”. Nine participants created a group with only Item 11; most of them named the group “Waiting time”. Item 11 constituting its own factor is in line with the Four-Factor model by Borsci & Schmettow (2023).

Figure 4

Clustered Heatmap of the Card Sorting



Note: Clustered heatmap of the combined Jaccard score of the open card sort. The numbered clusters, outlined in black show the Four-Factor Model that was created during the analysis. Two types of points of interest are outlined and labelled with letters. All cells within the models whose values are below the threshold of at least 40% agreement are outlined in light blue. Cells outside the factors that exceed the threshold of 40% are outlined in dark blue.

Four-factor model

The Four-Factor Model consists of the factors of User-friendliness (Items 1, 2, 11), Comprehension (Items 3, 4, 5, 6), Quality of information in answer (Items 7, 8, 9) and Privacy (Item 11). The model was extracted by applying a threshold of at least 40% agreement.

User-friendliness – Items 1, 2, 11

In the heatmap, the first extracted factor can be seen in the top left of the heatmap and is labelled “1”. It contains Items 1, 2 and 11. Item 1 was grouped 20 times with Item 11. Similarly, Item 2 was paired 20 times with Item 11 and 14 times with Item 3. Participants who created groups with the three items most commonly chose names related to the visibility of the chatbot. However, these only cover items 1 and 2 and are therefore unclear. Other names given

to the groups were “Usability”, “User experience” and “User-friendliness”. Out of those, the name "User-friendliness" was picked.

Comprehension- Items 3, 4, 5, 6

The factor labelled “2”, contains items 3, 4, 5 and 6. Item 3 was grouped with item 4 21 times (44.68 %), with item 5 32 times (68.09 %). Item 4 was grouped with item 6 28 times (59.57 %). Item 5 was grouped with item 4 19 times (40.43 %).

The point of interest A shows that item 6 was grouped with item 3 and item 5 less than 40% of the time. Item 6 was grouped with item 3 and item 5 36.17% of the time. Three participants created groups with the four items, and all of them used names that referred to comprehension or content. Therefore, the group was called “Comprehension”.

Table 2

Factor structure of model proposed by Borsci & Schmettow (2023) *and the new models generated with the heatmap*

Borsci- 4 factor	6 factor model	4 factor model
ACC (1, 2)	ACC (1,2)	FRI (1,2,11)
QUAL (3,4,5,6,7,8,9)	EXP (3,5) UND (4,6) INF (7,8,9)	COM (3,4,5,6) INF (7,8,9)
PRIV (10)		
TIME (11)	PRIV (10) TIME (11)	PRIV (10)

Confirmatory factor analysis

To check the construct validity of the Four-Factor Model and the factor models created with the help of the heatmap, a cfa is carried out.

Cronbach's alpha was calculated to assess the internal consistency of the scale. Cronbach alpha of 0.9 indicates that the scale is strongly reliable. The Shapiro-Wilk test returned significant ($p<.001$), indicating that the data are not normally distributed. A density plot of the total scores (see Appendix A) shows a bimodal distribution that appears slightly negatively skewed. The skewness of the BUS scores was -0.4, indicating that the scores are distributed fairly symmetrically. The kurtosis was found to be -0.7 indicating that the distribution was more platykurtic than a normal distribution.

The factor analysis results are summarized in Table 3 and an overview of the analysed models can be found in Table 2.

The results suggest that the Four-Factor Model proposed by Borsci & Schmettow (2023), was not well suited for the data as the RMSEA was 0.0771, which is above the 0.7 threshold. Nevertheless, in the study by Borsci and Schmettow (2023) with a larger sample, the Four-Factor model fulfilled the RMSEA threshold. Therefore, a more lenient threshold of 0.8 is accepted for the following analyses. According to MacCallum et al. (1999), the RMSEA of 0.8 indicates a mediocre fit.

With a more lenient RMSEA, the Four-Factor Model by Borsci is acceptable. The Four-Factor Model by Borsci & Schmettow (2023) has a Chi-square/df of 2.66, a CFI of 0.965 and an SRMR of 0.04. The factor loading of the Four-Factor all meet the minimum of 0.6. The items with the lowest factor loadings are Item 2 and Item 5, with a factor loading of 0.7 (see Appendix A).

For the Six-Factor Model, the RMSEA was 0.066. In the Six-Factor Model, Item 5 has the lowest factor loading with a value of 0.7. The Chi-square to df ratio is 2.20; the CFI equals 0.980 and the SRMR equals 0.029. Of all tested models (see Table 3), the Six-Factor Model had the lowest AIC value.

The hierarchical model has four first-order factors and three second-order factors. The first-order factors are the factors of the Four-Factor Model by Borsci. The factor Functional interactive conversation (QUAL) has three second-order factors. The three underlying factors are identical to three of the factors of the Six-Factor Model. The first underlying factor is Quality of Chatbot expressions (EXP) and contains items 3 and 5. The second factor is constituted by items 4 and 6 and is called Contextual understanding (UND). Lastly, the third factor Quality of information in answer (INF), contains items 7, 8 and 9. The factor analysis results showed that the RMSEA of the hierarchical factor model is 0.068, which is meeting the threshold of 0.7. The hierarchical model has Chi-square to df ratio of 2.304, a CFI of 0.975 and an SRMR of 0.038. The items with the lowest factor loadings are Item 2 and Item

5, with a factor loading of 0.7 (see Appendix A). Of all tested models, the hierarchical model had the lowest BIC and the second lowest AIC. Therefore, the hierarchical model appears to be the model that best describes the data in this analysis, as all the fit measurements met the thresholds and both AIC and BIC are comparably low.

The newly created Four-Factor Model cannot be considered acceptable. The factor loading of item 11 on the Factor "User-friendliness" was 0.3 and did not meet the threshold of 0.6. Further, the RMSEA was 0.079, thereby only fulfilling a more lenient threshold of 0.8. In the Four-Factor Model, Item 11 has the lowest factor loading with a value of 0.3. The Chi-square to df ratio is 2.762; the CFI is equal to 0.964 and an SRMR of 0.056.

Table 3

Comparative analysis of the fit indexes of the new models and the Four-Factor Model and Five-Factor Model

Fit indexes	Criteria	New models			Old models	
		6-Factor	hierarchical	4-Factor	4-Borsci	5-Borsci
Chi square/df	<3	2.203	2.304	2.762	2.664	2.648
CFI	>=0.9	0.980	0.975	0.964	0.965	0.969
RSMEA	<0.07	0.066	0.068	0.079	0.077	0.077
SRMR	<0.08	0.029	0.038	0.056	0.040	0.036
AIC	lowest	8400.858	8405.834	8424.275	8421.127	8417.889
BIC	lowest	8528.076	8511.243	8522.414	8515.632	8526.933

Discussion

The aim of the present study was to explore the mental models of participants for the items of the BUS-11 scale with card sorting. A heatmap was used to compare the mental model of the Four-Factor structure proposed by Borsci & Schmettow (2023). Further, three new factor models were generated by analysing a clustered heatmap of the item-to item level matrix. In order to test the construct validity of the three new factor models and the Four-Factor Model proposed by Borsci and Schmettow (2023) a confirmatory factor analysis was conducted.

Findings

The results of the heatmap analysis show that the Four-Factor Model proposed by Borsci & Schmettow can only be deducted from the item-item matrix when a lenient threshold of at least 30% grouping frequency was applied. During the confirmatory factor analysis, the Four-Factor Model (Borsci & Schmettow, 2023) did not fulfill the criteria for the RMSEA that were used in Borsci et al. (2022). The Four-Factor Model could only be accepted with a more lenient cut-off score, that indicates a mediocre fit. This indicates that the results of the factor analysis should be treated with caution, as prior analysis with a bigger dataset showed an acceptable fit (Borsci and Schmettow, 2023).

This result of the heatmap analysis suggests that the participants had rather differentiated models of communication. The heatmap analysis with a strict threshold of agreements lead to the identification of a Six-Factor Model. The difference between the Four-Factor model by Borsci and the Six-Factor model is, that the factor Functional interactive conversation is split into three factors. In this study, the factors were called Quality of Chatbot Expressions, Contextual understanding and Quality of Information in answers. During the confirmatory factor analysis, the model fit was acceptable. The factor analysis showed that a hierarchical model was fitting the used data the best. The hierarchical model used the factors of the Four-Factor Model by Borsci & Schmettow (2023) as first-order factors, and the three novel factors of the novel Six-Factor Model as second-order models. The main value of the hierarchical model is that it could identify issues with chatbots more precisely than prior models.

The heatmap analysis with a less restrictive threshold resulted in a new Four-Factor Model. Similar to the results of the exploratory factor analysis of the design-o-metric approach by Borsci & Schmettow (2023), the heatmap analysis also suggested that the Items

1, 2 and 11 can be combined into one factor. Nevertheless, the factor loadings were too low to be deemed acceptable in the study by Borsci and Schmettow (2023) as well as this one. This similarity in the findings is unexpected, since both studies chose very different approaches.

A potential reason why the items of the factor Accessibility and Responsiveness were frequently grouped together in this study and correlated in the study by Borsci & Schmettow (2023) is that individuals rely on their mental models to answer the items of the BUS-11. For the evaluation of the BUS-11 participants first completed task with a chatbot and evaluated it thereafter. As a result, participants had to memorise their experiences and then recall it to answer the items of the BUS.

Numerous studies have shown that mental models influence both the formation of memories and their retrieval (Jones et al., 2011). Participants could therefore remember their experiences with the chatbot in a simplified form. When they are then asked about the chatbot, they may not only evaluate the attribute in question, but also the category it belongs to in their mental model.

Further, Borsci and Schmettow (2023) noted, that the factors Accessability and Responsiveness both contain items that are not goal oriented. One reason, why both studies found a similar factor structure could be due to the fact, that the participants in both studies were familiarized with the chatbots in the same way. Performing a task in a study setting and evaluating the chatbot in detail may have caused the participant to select a goal-directed approach. Selective attention could thereby cause individuals to only focus on task-relevant information, while filtering out non-essential information (Bundesen, 1996 as cited by Dayan at al., 2000). This could facilitate participants to perceive non-functional items as insignificant, if no serious problems occur in these areas.

However, item10, which is concerned with privacy, was hardly grouped with the other non-functional aspects, and did not correlate significantly with other items in the study by Borsci and Schmettow (2023). A potential explanation could be that participants understood their privacy security relevant. Studies have shown that people exhibit perceptual attention biases towards threatening appearing stimuli, even when performing tasks that do not require it.

Limitation

The results of the card sorting showed that some participants misinterpreted the task. In addition, some participants only took part in the first part of the study. This could indicate that the participants were already insufficiently concentrated by that point. In addition, participants had to switch platforms for the second part of the study, which could have failed because of technical issues. Further, participant feedback occasionally reflected dissatisfaction with the chatbot task, as some participants had difficulty finding the chatbot itself or its translation function. The chatbot was originally chosen because it supports all three languages used, ensuring that the experimental set-up is similar for all participants. However, in future studies, it should be ensured that tasks that serve to familiarize participants with the topic do not trigger frustration and thus do not consume too many cognitive resources

A weakness of the analysis of group names is, that the determination of whether names are related depends on the researcher's judgement. An alternative, more sophisticated method of analysis would be to use natural language processing techniques to evaluate the similarity of group names. This could involve algorithms such as word2vec, which uses a neural network model to learn word associations from text. One of the challenges here is to train the models on phrases rather than single words. To evaluate whether semantically similar groups also contain similar items, the semantic similarity could be represented by multidimensional scaling; this representation could be compared with a representation of multidimensionally scaled representation of the groupings as demonstrated by Paea et al (2022).

Lastly, the results of the factor analysis should be treated with caution. In previous studies with bigger datasets, both the Four-Factor and Five-Factor models showed an acceptable fit (Borsci & Schmettow, 2023). These results could not be replicated in this study. One potential reason is that the used dataset with 280 observations was too small, as bigger datasets tend to lead to more precise estimations during factor analysis (MacCallum & Widaman, 1999). In addition, only 10 different chatbots were evaluated in the data set. This should be viewed critically, as a small number of chatbots will most likely not reflect the attributes of the chatbot population. In this context, it has to be considered that the BUS-11 aims to evaluate chatbots while psychometrics is commonly used to measure the characteristics of participants. Thus, the dataset should ideally include the BUS-11 scores of a large sample of chatbots. This measure prevents a lack of correspondence between the chatbot characteristics in the sample and the population of chatbots. That the used dataset

showed a non-normal distribution of the data also speaks for an insufficient amount of evaluated chatbots.

Conclusion

The aim of the present study was to utilise card sorting to explore the mental models of participants for the Bot Usability Scale developed by Borsci et al. (2022). A heatmap was used to compare the mental models of participants to the Four-Factor Structure proposed by Borsci & Schmettow (2023). The analysis showed that the Four-Factor Model could be extracted from the heatmap, when a relatively low threshold of 30% participant agreement was applied. This was because the participants had sorted the items of the factor Functional interactive conversation into more differentiated groups. However, the analysis of the face validity showed that the participants who had formed identical groups to the Four-Factor Model also used similar names to the factor names.

Further, three new factor models were generated during the analysis of the participants' mental models. During the confirmatory factor analysis, the newly constructed hierarchical model with four first-order and three second-order factors was found to have the best fit. This model could potentially diagnose problems more accurately than the Four-Factor Model proposed by Borsci & Schmettow (2023).

In future research, the method to familiarise the participant with the chatbot could be altered. The previously chosen approach could be eliciting the observer bias, causing participants to be more attentive to goal-oriented aspects of use and thereby influencing their ratings of the BUS-11 items.

References

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

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). https://doi.org/10.2196/20404

Binkhorst, V., Fiebig, T., Krombholz, K., Pieters, W., & Labunets, K. (2022). Security at the End of the Tunnel: The Anatomy of VPN Mental Models Among Experts and Non-Experts in a Corporate Context.

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

Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & van der Velde, F. (2022). A confirmatory factorial analysis of the chatbot usability scale: A multilanguage validation. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-022-01690-0

Borsci, Schmettow (2023) The universality of the chatBot Usability Scale (BUS-11): from a psychometric to a designometric perspective. Manuscript in preparation.

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike Information Criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, *11*(3). https://doi.org/10.1002/wics.1460

De Keyser, A., Köcher, S., Alkire (née Nasr), L., Verbeeck, C., & Kandampully, J. (2019). Frontline Service Technology Infusion: Conceptual Archetypes and future research

directions. *Journal of Service Management*, *30*(1), 156–183. https://doi.org/10.1108/josm-03-2018-0082

Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with Chatbots: Findings from a questionnaire study. *Quality and User Experience*, *5*(1). https://doi.org/10.1007/s41233-020-00033-2

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? an exploratory interview study. *Internet Science*, 194–208. https://doi.org/10.1007/978-3-030-01437-7_16

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

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Hatem, G., Zeidan, J., Goossens, M., & Moreira, C. (2022). Normality testing methods and the importance of skewness and kurtosis in statistical analysis. *BAU Journal - Science and Technology*, *3*(2). https://doi.org/10.54729/ktpe9512

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and Research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

ISO (2018). Ergonomics of human system interaction-Part 11: Usability:Definition and concepts (9241). Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Jones, N. A., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society*, *16*(1). https://doi.org/10.5751/es-03802-160146

Lester, J. C., Branting, K., & Mott, B. W. (2004). Conversational agents. In *The Practical Handbook of Internet Computing* (pp. 1–17). CRC Press.

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, Sus, and umux. *International Journal of Human–Computer Interaction, 34*(12), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805

Li, CH. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res 48*, 936–949. https://doi-org.ezproxy2.utwente.nl/10.3758/s13428-015-0619-7

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p16-1094

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989x.4.1.84

Nielsen, J. (2012). *Usability 101: Introduction to usability*. Nielsen Norman Group. Retrieved October 10, 2022, from https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Paea, S., Katsanos, C., & Bulivou, G. (2021). Information architecture: Using K-means clustering and the best merge method for Open Card Sorting Data Analysis. *Interacting with Computers*, *33*(6), 670–689. https://doi.org/10.1093/iwc/iwac022

Radziwill, N. M., & Benton, M. C. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents. *Software Quality Professional*, *19*(3), 25-36. https://doi.org/10.48550/arXiv.1704.04579

Rese, A., Ganster, L. & Baier, D. (2020, September). Chatbots in retailers' customer communication: How to measure their acceptance? *Journal of Retailing and Consumer Services*, *56*, 102176. https://doi.org/10.1016/j.jretconser.2020.102176

Rich, A., & McGee, M. (2004). Expected usability magnitude estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(5), 912–916. https://doi.org/10.1177/154193120404800536

Schmettow M (2021). New statistics for design researchers. Springer International Publishing, Cham

Schmettow, M., & Sommer, J. (2016). Linking card sorting to browsing performance – are congruent municipal websites more efficient to use? *Behaviour & Information Technology*, *35*(6), 452–470.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4–11. https://doi.org/10.12691/ajams-9-1-2

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a Research. *SSRN Electronic Journal*, *5*(3), 28–36. https://doi.org/10.2139/ssrn.3205040

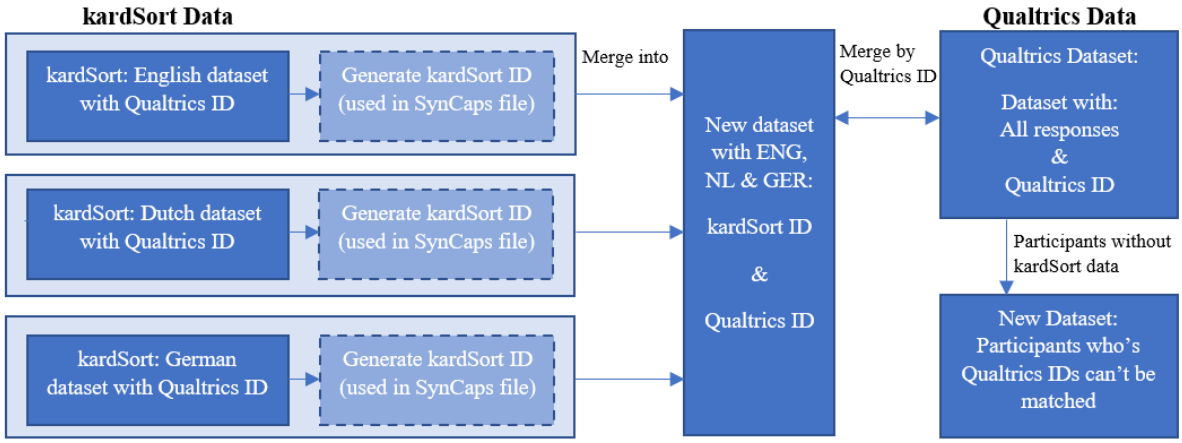
Urdan, T. C. (2022). *Statistics in plain english*. Routledge.

Vinyals, O., & Le, Q. (2015). A Neural Conversational Model arXiv preprint arXiv:1506.05869, 2015.

Xu, Y., Zhang, J., & Deng, G. (2022, July 22). Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.902782

Appendix A: Tables and Figures

Functionality of the R-script to merge the datasets from Qualtrics and kardSort



Matrix with the frequency of item groupings

	Item 1	Item 2	Item 11	Item 7	Item 8	Item 9	Item 3	Item 5	Item 4	Item 6	Item 10
Item 1	47	45	20	3	4	2	16	11	6	4	4
Item 2	45	47	20	1	4	3	14	11	4	5	3
Item 11	20	20	47	4	8	9	17	12	13	7	5
Item 7	3	1	4	47	41	26	10	15	15	17	11
Item 8	4	4	8	41	47	25	8	16	16	14	8
Item 9	2	3	9	26	25	47	13	22	19	22	9
Item 3	16	14	17	10	8	13	47	32	21	17	8
Item 5	11	11	12	15	16	22	32	47	19	17	8
Item 4	6	4	13	15	16	19	21	19	47	28	9
Item 6	4	5	7	17	14	22	17	17	28	47	9
Item 10	4	3	5	11	8	9	8	8	9	9	47

Tables of the names of groups identical to the Factors of the Borsci Four-Factor Model

	Participant groups identical to Borsci Factors			
Lang	ACC (1,2)	QUAL (3,4,5,6,7,8,9)	PRIV (10)	TIME (11)
EN	- Accessibility (3x) - Ease of finding - Findability - Finding chatbot - Finding the Chatbot - finding chatbot function - Chat box findability - Locating Ease - Location of Chatbot-icon - Detection - Visualization - User interaction	- Quality of the answers of the chatbot - information provided	- Privacy (6x) - privacy issues - privacy and safety - Data privacy - transparency	- Time (2x) - Waiting time (5x) - Performance
DE			- Datenschutz (2x)	- Vorteil
NL	- Vindbaarheid		- Privacy (4x) - Zorgen	

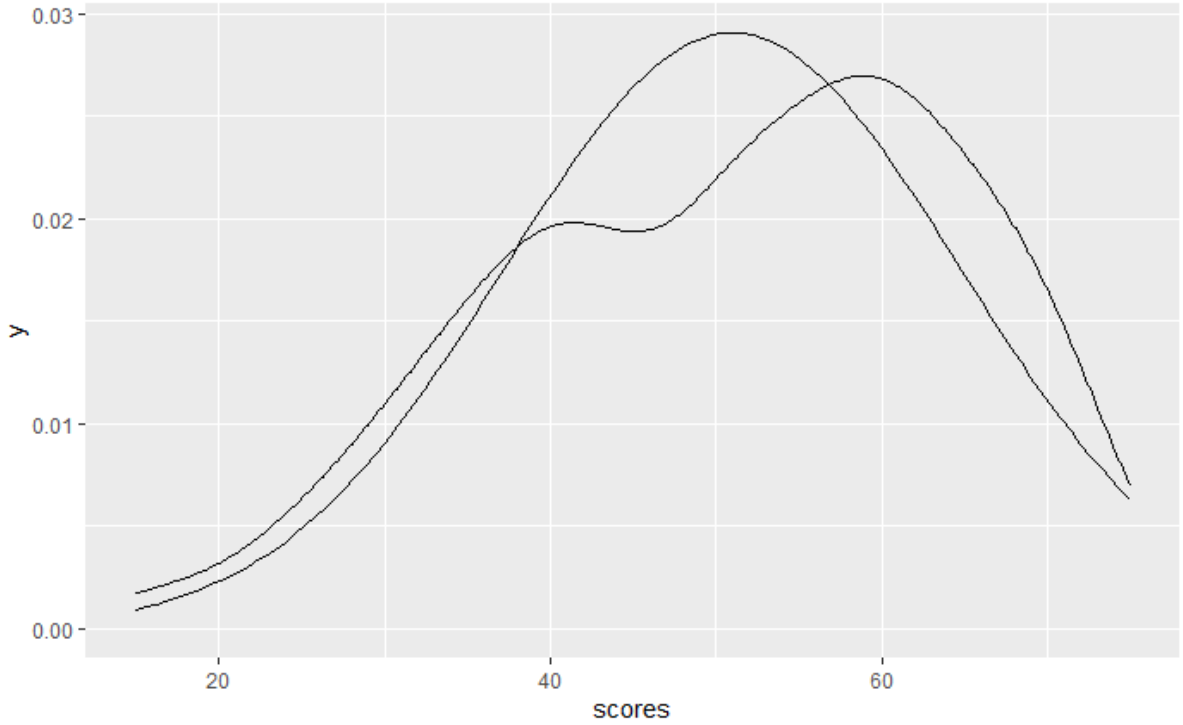
Tables of the names of groups identical to the Factors found during Heatmap analysis- Six-Factor model

Lang	Participant groups identical to Heatmap Factors		
	EXP (3,5)	UND (4,6)	INF (7,8,9)
EN	- Communication - Quality of chatbots expression	- Chatbot understand context - Contextual understanding - usability	- Information - Simplicity and accuracy - Given information - Quality of information in answer
DE	- Verständlichkeit der Antworten		- Information
NL	- Communicatie	- context	- Inhoud

Tables of the names of groups identical to the Factors found during Heatmap analysis- Four-Factor model

Lang	Participant groups identical to Heatmap Factors	
	FRI (1,2,11)	COM (3,4,5,6)
EN	- Interface - Usability	- Comprehension
DE	- User Experience - Sichtbarkeit	- Inhalt / Verständnis - Verständlichkeit
NL	- Chatbot ontdekken - Gebruiksvriendelijkheid	

Density plot of the total scores of the BUS-11



Factor loading tables: Confirmatory factor analysis

Borsci-Four-Factor

	ACC	QUAL	PRIV	TIME
Item 1	1	0	0	0
Item 2	0.7	0	0	0
Item 3	0	0.9	0	0
Item 4	0	0.8	0	0
Item 5	0	0.7	0	0
Item 6	0	0.9	0	0
Item 7	0	0.8	0	0
Item 8	0	0.8	0	0
Item 9	0	0.9	0	0
Item 10	0	0	1	0
Item 11	0	0	0	1

Newly created four Factor Model

	FRI	EXP	INF	PRIV
Item 1	0.9	0	0	0
Item 2	0.8	0	0	0
Item 11	0.3	0	0	0
Item 3	0	0.9	0	0
Item 5	0	0.7	0	0
Item 4	0	0.8	0	0
Item 6	0	0.9	0	0
Item 7	0	0	0.8	0
Item 8	0	0	0.8	0
Item 9	0	0	0.9	0
Item 10	0	0	0	1

Newly created Six-Factor Model

	ACC	EXP	UND	INF	PRIV	TIME
Item 1	1	0	0	0	0	0
Item 2	0.8	0	0	0	0	0
Item 3	0	0.9	0	0	0	0
Item 5	0	0.7	0	0	0	0
Item 4	0	0	0.8	0	0	0
Item 6	0	0	0.9	0	0	0
Item 7	0	0	0	0.8	0	0
Item 8	0	0	0	0.8	0	0
Item 9	0	0	0	0.9	0	0
Item 10	0	0	0	0	1	0
Item 11	0	0	0	0	0	1

Newly created Hierarchical Factor Model

	ACC	EXP	UND	INF	PRIV	TIME	QUAL
Item_1	1.0	0.0	0.0	0.0	0	0	0
Item_2	0.7	0.0	0.0	0.0	0	0	0
Item_3	0.0	0.9	0.0	0.0	0	0	0
Item_5	0.0	0.7	0.0	0.0	0	0	0
Item_4	0.0	0.0	0.8	0.0	0	0	0
Item_6	0.0	0.0	0.9	0.0	0	0	0
Item_7	0.0	0.0	0.0	0.8	0	0	0
Item_8	0.0	0.0	0.0	0.8	0	0	0
Item_9	0.0	0.0	0.0	0.9	0	0	0
Item_10	0.0	0.0	0.0	0.0	1	0	0
Item_11	0.0	0.0	0.0	0.0	0	1	0

Appendix B: Qualtrics survey

Open card sorting

Start of Block: Introduction

Dear participant, Thank you for participating in this research! **Please use a PC or laptop to participate in this study!** Before we begin, you will receive information about the research and your rights. Taking part in this research is voluntary, and you can withdraw at any moment. Withdrawing will not have negative consequences for you.

Purpose of the research
This research is about investigating how humans experience interaction with different chatbots. Chatbots are programmes that communicate with you like a human but do not require a human to operate.

Content of the research
Taking part in this research consists of different components. After reading this introduction, we will ask you to give informed consent and inquire about your demographics. Further, we will ask you to interact with a chatbot and evaluate your experience. Lastly, you will be redirected to a different website to sort cards.

Data processing
The data of this research will be used to gather the mental models of our participants. The information will be used for a Bachelor Thesis. Your data will be anonymised and cannot be traced back to you. Your data will not be shared with third parties. The anonymised information is stored in a secure environment and kept for use in future studies. This research is approved by the Ethics Committee of the University of Twente. You can navigate this survey by clicking the arrows at the bottom on the page. To proceed, please click on the arrow on the bottom right.

End of Block: Introduction

Start of Block: Informed Consent

Consent Form

Taking part in the study
I have read and understood the study information. I voluntarily take part in this research and understand that I can refuse to answer questions. I know that I can withdraw from this study at any time, without having to give a reason. I understand that I have to interact with a chatbot and that participating does not involve any risks. I am at least 18 years old.
Use of the information in the study
I understand that providing demographic data, interacting with a chatbot, and filling in a questionnaire after is also part of the study. Further, I will be asked to sort cards in the final part of the study

Future use and reuse of the information by others
I understand that the information that I provide will be used for a bachelor thesis. I know that all information will be anonymised and stored in a secure environment. I consent that the anonymised information provided by me is kept for use in future studies.
Contact information for questions about Your rights as a participant
If you have any questions regarding your participation in this study, you can email c.e.wermter@student.utwente.nl. You can also reach the supervisor by emailing j.landwehr@utwente.nl. If you have any questions about your rights as a participant, the use of your data, or other questions and concerns about this research, you can contact the secretariat of the Ethics Committee of the Faculty of Behavioural, Management, and Social Sciences of the University of Twente: ethicscommittee-bms@utwente.nl. Do you consent to participating in this research?

Yes, I consent (1)

No, I want to stop (2)

End of Block: Informed Consent

Start of Block: Demographic data

Demographic Data 1/2

You will now receive questions about your demographics.

How old are you in years?

_____

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

☐ Man (1)

☐ Woman (2)

☐ Female-to-Male (FTM)/Transgender Male/Trans Man (3)

☐ Male-to-Female (MtF)/Transgender Female/Trans Woman (4)

☐ Genderqueer, neither exclusively male or female; (5)

☐ Additional Gender Category/(or Other), please specify (6)
_____

☐ Decline to answer (7)

What is your sex (as assigned at birth)?

Male (1)

Female (2)

Intersex (3)

What is your nationality?

Dutch (1)

German (2)

Other: (3) _____

What is your level of English proficiency?

No proficiency- Knowing few to no words; unable to form full sentences. (1)

Elementary proficiency-Able to form basic sentences and answer simple questions. (2)

Limited proficiency- Able to use social phrases and carry limited casual conversations. (3)

Basic proficiency- Having fairly extensive vocabulary and being able to hold conversations. (4)

Full proficiency- Able to have advanced discussions on a wide range of topics. (5)

Native/bilingual proficiency- Able to speak completely fluent. (6)

Are you diagnosed with ADHD or ADD?

Yes (1)

No (2)

Are you medicated for ADHD/ADD?

Yes (1)

No (2)

End of Block: Demographic data

Start of Block: Experience

Demographic Data 2/2

You will now receive a number of statements about your experience with chatbots. Please indicate for each statement how much you agree.

Familiarity

	Fully Disgree (1)	Disagree (2)	Neutral (3)	Agree (4)	Fully Agree (5)
I am familiar with chatbots and/or other conversational interfaces. (1)					
I know how chatbots work. (2)					
I feel confident with using chatbots. (3)					

Before we begin, please indicate how often you use chatbots.

	Never (6)	Seldom (5)	1 time per month (4)	2-3 times a month (3)	4-6 times a month (2)	Daily (1)
How many times do you use a chatbot per months? (1)						

End of Block: Experience

Start of Block: Task Explanation

Explanation Task

We will now ask you to interact with a chatbot. You will receive a link to a website. Please open this website in a second tab. We will then give you a task to solve with the help of the chatbot.

Do not worry if you cannot solve the task. You can simply continue to participate and complete the questionnaire about the chatbot. If you solve the task, please also continue to answer the questionnaire about the chatbot.

Note: You never have to provide any personal information.

End of Block: Task Explanation

Start of Block: Chatbot #1

Chatbot

Please open the link in a second tab (leave the survey tab open) and find the chatbot. In the next step, you will be asked to solve a task with the chatbot.

To change the chatbot to English, click on "Translate" and select English.

https://www.oxxio.nl/klantenservice

Perform the following task using the chatbot:

What are the advantages of a Smart meter? (two answers)

- ☐ You can view your energy consumption anytime. (1)
- ☐ You can control your home remotely. (2)
- ☐ You do not have to report your meter readings manually. (3)
- ☐ You can automatise processes such as temperature regulation. (4)

Was it possible to complete the task?

Yes (1)

No (Why not, please specify) (2) _____

I am not sure (3)

Please answer the following questions on the basis of your experience interacting with the chatbot.

	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
The chatbot gives me the appropriate amount of information. (7)					
Communicating with the chatbot was clear. (3)					
The chatbot function was easily detectable. (1)					
The chatbot was able to keep track of context. (4)					
I find that the chatbot understands what I want and helps me achieve my goal. (6)					
It was easy to find the chatbot. (2)					
The chatbot's responses were easy to understand. (5)					
My waiting time for a response from the chatbot was short. (11)					
The chatbot gives me the information I need. (8)					
I believe the chatbot informs me of any possible privacy issues. (10)					

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

End of Block: Chatbot #1

Instructions

Introduction:

Now we would like you to sort cards. The 11 cards that we will show you belong to a scale that assesses the user satisfaction with chatbots. We are interested to know in which groups you would place these items. We would like you to group items together that you think belong together.

Instructions:

Please, group cards together that you believe belong in the same category. You can do this by creating a category and then dragging items into its field. Please give each category a name that you think describes it. You can edit the category name and change the location of the items if needed.

As an example:

- *the items apple, cheese, pear, yogurt and salami are given*
- *one might group them as: milk products (cheese, yogurt); fruit (apple, pear); meat products (salami)*

There is no right or wrong way to sort the cards. It's important that you sort the cards into categories that are meaningful for you.

Note:

- **One card alone can be one category**
- **Please, sort all the cards into categories**

Appendix C: Analysis script

````
---
title: "Heatmap"
author: "Chiara Wermter"
date: "2022-12-03"
output:
  word_document: default
  html_document: default
  pdf_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r load libraries}
#function to load packages that are required and not installed

if (!require("sass")) {
  install.packages("sass", dependencies = TRUE)
  library(sass)
}
if (!require("tidyr")) {
  install.packages("tidyr", dependencies = TRUE)
  library(tidyr)
}
if (!require("tidyverse")) {
  install.packages("tidyverse", dependencies = TRUE)
  library(tidyverse)
}
if (!require("dplyr")) {
  install.packages("dplyr", dependencies = TRUE)
  library(dplyr)
}
if (!require("ggplot2")) {
  install.packages("ggplot2", dependencies = TRUE)
  library(ggplot2)
}
if (!require("gplots")) {
  install.packages("gplots", dependencies = TRUE)
  library(gplots)
}
if (!require("lattice")) {
  install.packages("lattice", dependencies = TRUE)
  library(lattice)
}
if (!require("tibble")) {
  install.packages("tibble", dependencies = TRUE)
  library(tibble)
````
````

```
}
#load packages
if (!require("rlang")) {
  install.packages("rlang", dependencies = TRUE)
  library(rlang)
}
if (!require("BiocManager")) {
  install.packages("BiocManager", dependencies = TRUE)
  library(BiocManager)
}
if (!require("ComplexHeatmap")) {
  install.packages("ComplexHeatmap", dependencies = TRUE)
  library(ComplexHeatmap)
}
if (!require("vctrs")) {
  install.packages("vctrs", repos = "https://packagemanager.rstudio.com/cran/latest")
  library(vctrs)
}
if (!require("haven")) {
  install.packages("haven", dependencies = TRUE)
  library(haven)
}
```

##Input
In this section the script will ask for the sample size and language requirements.
The entered values will be used throughout the analysis. Thus, the user doesn't
need to change individual lines of the script, when a dataset with a different sample
size is loaded or the language requirements are changed.

At this point, re-run this chunk if you change the variables.

Further, you can set the colours for the heatmap in this section.

```
    #table with language requirements
    1 - No proficiency- Knowing few to no words; unable to form full sentences.
    2 - Elementary proficiency-Able to form basic sentences and answer simple questions.
    3 - Limited proficiency- Able to use social phrases and carry limited casual conversations.
    4 - Basic proficiency- Having fairly extensive vocabulary and being able to hold
conversations.
    5 - Full proficiency- Able to have advanced discussions on a wide range of topics.
    6 - Native/bilingual proficiency- Able to speak completely fluent.

```{r input and variables}
##input
sample.size <- readline("Please enter sample size!") #47
sample.size <- as.numeric(sample.size)
```
```

```
lang_req <- readline("Please enter minimum language requirement!") #4
lang_req<-  as.numeric(lang_req)

heatc= c('#fcfc05','#fc9905','#db3d0d') #define colours of heatmap
```

Combining Qualtrics and kardSort Datasets
In the following two sections the the kardSort and Qualtrics dataset are combined, allowing to view the Qualtrics and kardSort ID as well as the language in which the participant finished
the card sorting.
You can use this script to prepare the Syncaps file, by identifying participants that did not finish parts of the study or did not meet the language requirements.

Further, you can group participants by a set of responses to ease creating grouped Syncaps files.

Please enter the data path of the indicated files to run the script correctly.

```{r load Qualtrics & kardSort datasets}
#read english questionaire responses (kardSort)
q_eng<-read.csv("C:\\Users\\Alison Wang\\Desktop\\Bachelor\\analysis_scripts\\files\\eng_qrespon.csv")

#read dutch questionaire responses (kardSort)
q_dut<-read.csv("C:\\Users\\Alison Wang\\Desktop\\Bachelor\\analysis_scripts\\files\\dut_qrespon.csv")

#read german questionaire responses (kardSort)
q_ger<-read.csv("C:\\Users\\Alison Wang\\Desktop\\Bachelor\\analysis_scripts\\files\\ger_qrespon.csv")

#read qualtrics questionaire responses (Qualtrics)
qualtr <- read_sav("C:/Users/Alison Wang/Desktop/Bachelor/analysis_scripts/files/qualtrics_respon.sav")
```

```{r Merging Qualtrics and kardSort dataset}

####### English
q_eng <- tibble::rowid_to_column(q_eng, "kID")          #add kartSort participant ID

####### Dutch
q_dut <- tibble::rowid_to_column(q_dut, "kID")          #add kartSort participant ID

####### German
q_ger <- tibble::rowid_to_column(q_ger, "kID")          #add kartSort participant ID

######combine all languages into one dataframe
ID_all <- rbind(q_eng, q_dut, q_ger)
```

```
#clean up
rm(q_eng, q_dut, q_ger) #remove obsolete dataframes

ID_all<-ID_all[-c(2:4)]                          #isolating id provided by participant (in answer-column)
colnames(ID_all)[colnames(ID_all) == "answer"] <- "ID"  #renaming column from "answer" to "ID" (qualtricsID)
ID_all$ID <- gsub("[^[:digit:] ]", "", ID_all$ID)       #removing everything except for numbers from ID

######qualtrics
#clean up
qualtr<-qualtr[-c(1:18,20:27,57)]                    #delete clutter
qualtr <-qualtr %>%
  tidyr::drop_na(nat) #drop columns with NA at required questions

qualtr <- qualtr %>%                                 #move columns Id and language to the front
  dplyr::select(ID, Q_Language,everything())

qua.ref <- qualtr                                   #create reference df for codebook/
sjPlot::view_df(qua.ref)                             #show codebook

qualtr <- qualtr %>%                                 #transform the data type of the columns to character
  dplyr::mutate(across(everything(), as.character))


#recode variables
qualtr$sex<- dplyr::recode(qualtr$sex, '1'="male", '2'="female",'3'="intersex")
qualtr$nat<- dplyr::recode(qualtr$nat, '1'="Dutch", '2'="German")       #replace coded values with text
qualtr[qualtr$nat=="3", "nat"]<- qualtr[qualtr$nat=="3", "nat_3_TEXT"]  #when participants selected "other" -> insert their text entry from the next column
qualtr$nat<- sub("^$", "Other", qualtr$nat)                          #when participant selected "other" and didnt enter text -> replace the blank with "other"
qualtr$nat_3_TEXT <- NULL #drop obsolete column

#combine multiple choice options from TaskQuestion into binary
qualtr$CH_TaskQuestion_1<- case_when (qualtr$CH_TaskQuestion_1 =="1" & is.na(qualtr$CH_TaskQuestion_2) & qualtr$CH_TaskQuestion_3 =="1" & is.na(qualtr$CH_TaskQuestion_4)            ~ "1",  qualtr$CH_TaskQuestion_2 =="1" | qualtr$CH_TaskQuestion_4 =="1" | is.na(qualtr$CH_TaskQuestion_1) | is.na(qualtr$CH_TaskQuestion_3)            ~ "0")
qualtr$CH_TaskQuestion_2 <- NULL
qualtr$CH_TaskQuestion_3 <- NULL
qualtr$CH_TaskQuestion_4 <- NULL
colnames(qualtr)[colnames(qualtr) == "CH_TaskQuestion_1"] <- "TaskCb" #renaming column from "CH_TaskQuestion_1" to "TaskCb"
colnames(qualtr)[colnames(qualtr) == "CH_Complete"] <- "TaskCb_comp"
```

```
colnames(qualtr)[colnames(qualtr) == "CH_Complete_2_TEXT"] <- "TaskCb_issue"

colnames(qualtr)[colnames(qualtr) == "CH_BUS_1"] ="Item 1"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_2"] ="Item 2"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_3"] ="Item 3"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_4"] ="Item 4"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_5"] ="Item 5"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_6"] ="Item 6"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_7"] ="Item 7"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_8"] ="Item 8"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_9"] ="Item 9"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_10"] ="Item 10"
colnames(qualtr)[colnames(qualtr) == "CH_BUS_11"] ="Item 11"

#merge qualtrics dataframe with ID dataframe
comb<-base::merge(ID_all, qualtr, by="ID", all=TRUE) #this function is automatically dropping participants that didnt parttake in the second study part and those who have been deleted from qualtrics

Excluded <- subset(comb, is.na(kID)|is.na(age))
comb<-comb[!(is.na(comb$kID)|is.na(comb$age)),]

Insuf<- subset(comb,lang_prof < lang_req)  #subset responses to be deleted for documentation (insufficient language skills)
comb<-comb[!(comb$lang_prof < lang_req),] #delete participants from qualtr based on language requirements
rm(qualtr)
rm(ID_all)

```
```{r Participant characteristics}
comb$age <- as.numeric(comb$age)
mean(comb$age)
sd(comb$age)
min(comb$age)
max(comb$age)

summary(comb$nat)
table(comb$nat)
table(comb$sex)
```
```

## Heatmap
In this part of the script the heatmap is created.
PLEASE NOTE that the values are tranfered into percentages using the sample size that the user specified in the beginning.
If your sample size changed please re-run the chunk "r input and variables" to update the sample size.
```

```{r read dataset from Syncaps output + functions}
library(readxl)
cc_1M <- read_excel("C:/Users/Alison Wang/Desktop/Bachelor/data/cc_1M.xlsx")

create_numerical_matrix <- function(matrix_name, df){ ###function to create numerical matrices from df (for heat maps)
matrix_name <-as.matrix(df)
dims <- dim(matrix_name)
matrix_name <- as.numeric(matrix_name)
dim(matrix_name) <- dims
rownames(matrix_name) <-paste0("Item " , reference$ItemNo)
colnames(matrix_name) <-paste0("Item ", reference$ItemNo)
return(matrix_name)
}


```

```{r prepare dataset, echo=FALSE}

# Syntax rename with condition
colnames(cc_1M)[colnames(cc_1M) == "The chatbot function was easily detectable."] ="1"
colnames(cc_1M)[colnames(cc_1M) == "It was easy to find the chatbot."] ="2"
colnames(cc_1M)[colnames(cc_1M) == "Communicating with the chatbot was clear."] ="3"
colnames(cc_1M)[colnames(cc_1M) == "The chatbot was able to keep track of context."] ="4"
colnames(cc_1M)[colnames(cc_1M) == "The chatbot's responses were easy to understand."] ="5"
colnames(cc_1M)[colnames(cc_1M) == "I find that the chatbot undertands what I want and helps me achieve my goal."] ="6"
colnames(cc_1M)[colnames(cc_1M) == "The chatbot gives me the appropriate amount of information."] ="7"
colnames(cc_1M)[colnames(cc_1M) == "The chatbot only gives me the information I need."] ="8"
colnames(cc_1M)[colnames(cc_1M) == "I feel the chatbots reponses were accurate."] ="9"
colnames(cc_1M)[colnames(cc_1M) == "I believe the chatbot informs me of any possible privacy issues."] ="10"
colnames(cc_1M)[colnames(cc_1M) == "My waiting time for a response from the chatbot was short."] ="11"


#deleting 2nd column (Item description)
cc_1M$...2 <- NULL

#replace NA with participant number
cc_1M[is.na(cc_1M)]<-as.numeric(sample.size)

reference <- data.frame(cc_1M)
```

```
cc_1M$ItemNo <- NULL


```
##matrixes for heatmap
```{r create matrix in percentage}
cc_1M <- as.data.frame(sapply(cc_1M, as.numeric)) #transform dataframe to numeric
cc_1M <- cc_1M[,1:ncol(cc_1M)]/as.numeric(sample.size) #transforming to p
cc_1M <- cc_1M[,1:ncol(cc_1M)]*100 #transform into percentage

heatdata_freq <-create_numerical_matrix(heatdata, cc_1M)
heatdata <-create_numerical_matrix(heatdata, reference[c(2:12)])

```


##create and export ordered matrixes for different models
```{r}
if (!require("openxlsx")) {
  install.packages("openxlsx", dependencies = TRUE)
  library("openxlsx")
}
matrix_borsci<- create_numerical_matrix(matrix_borsci, reference[c(2:12)])
matrix_borsci_freq<- create_numerical_matrix(matrix_borsci_freq, cc_1M)

order.borsci <- c("Item 1","Item 2","Item 3","Item 4","Item 5","Item 6","Item 7", "Item
8","Item 9","Item 10","Item 11")
matrix_borsci <-matrix_borsci[ , order.borsci]
matrix_borsci <-matrix_borsci [order.borsci,]

matrix_borsci_freq <-matrix_borsci_freq[ , order.borsci]
matrix_borsci_freq <-matrix_borsci_freq [order.borsci,]

write.xlsx(as.data.frame(matrix_borsci), file="borsci.xlsx")

matrix_new<- create_numerical_matrix(matrix_new, reference[c(2:12)])
order.new <- c("Item 1","Item 2", "Item 11", "Item 3","Item 5","Item 4","Item 6","Item
9","Item 7", "Item 8", "Item 10")
matrix_new <-matrix_new[ , order.new]
matrix_new <-matrix_new [order.new,]
write.xlsx(as.data.frame(matrix_new), file="new.xlsx")
```


##heatmap
```{r create heatmap}
heat_own <-Heatmap(heatdata_freq,
       name = "Frequency of groupings in %",
       show_row_dend =FALSE,
       show_column_dend = FALSE,
       col= heatc,
```

```
        cluster_rows = TRUE,   # turn on clustering
        cluster_columns = TRUE,
        column_title = "Card Sort Heatmap - Clusteredi",
        width = ncol(heatdata_freq)*unit(8, "mm"),
        height = nrow(heatdata_freq)*unit(7, "mm"))

plot(heat_own)

heat_borsci <-Heatmap(matrix_borsci_freq,
        name = "Frequency of groupings in %",
        col= heatc,
        cluster_rows = FALSE,   # turn on clustering
        cluster_columns = FALSE,
        column_title = "Heatmap - Unclustered",
        width = ncol(heatdata_freq)*unit(8, "mm"),
        height = nrow(heatdata_freq)*unit(7, "mm"))
plot(heat_borsci)

heat_new <-Heatmap(matrix_new,
        name = "Frequency of groupings",
        col= heatc,
        cluster_rows = FALSE,   # turn on clustering
        cluster_columns = FALSE,
        column_title = "Heatmap - Clustered and ordered",
        width = ncol(heatdata_freq)*unit(8, "mm"),
        height = nrow(heatdata_freq)*unit(7, "mm"))
plot(heat_new)

#?Heatmap


```
```{r create an excelfile with numerical matrix with column order identical to heatmap clustering}
matrix_ownmodels <-create_numerical_matrix(matrix_ownmodels, reference[c(2:12)])
matrix_ownmodels_freq<- create_numerical_matrix(cc_1M, cc_1M)

order.own <- heat_own@column_names_param[["labels"]] ###grap order of items from heatmap
matrix_ownmodels <-matrix_ownmodels[ , order.own]##sorts according to clusters from heatmap
matrix_ownmodels <-matrix_ownmodels [order.own,]
write.xlsx(as.data.frame(matrix_ownmodels), file="own.xlsx")
```



```{r create correlation matrix}

#read prior BUS-11 results
```

```
CA_CFA_sav<-read_sav("C:\\Users\\Alison
Wang\\Desktop\\Bachelor\\analysis_scripts\\files\\CA_CFA_sav.sav")

#extract survey data for cfa from old data
cfa_data <- CA_CFA_sav %>%
  dplyr::select (12:14,17,20:26)

rm(CA_CFA_sav) # remove obsolete dataframe

#rename variables according to BUS-11 (from BUS-15)
colnames(cfa_data)[colnames(cfa_data) == "S1"] ="Item_1"
colnames(cfa_data)[colnames(cfa_data) == "S2"] ="Item_2"
colnames(cfa_data)[colnames(cfa_data) == "S3"] ="Item_3"
colnames(cfa_data)[colnames(cfa_data) == "S6"] ="Item_4"
colnames(cfa_data)[colnames(cfa_data) == "S9"] ="Item_5"
colnames(cfa_data)[colnames(cfa_data) == "S10"] ="Item_6"
colnames(cfa_data)[colnames(cfa_data) == "S11"] ="Item_7"
colnames(cfa_data)[colnames(cfa_data) == "S12"] ="Item_8"
colnames(cfa_data)[colnames(cfa_data) == "S13"] ="Item_9"
colnames(cfa_data)[colnames(cfa_data) == "S14"] ="Item_10"
colnames(cfa_data)[colnames(cfa_data) == "S15"] ="Item_11"

#create correlation matrix for scores
corr<-round(cor(cfa_data),2)

#plot correlation matrix as heatmap (for fun)
BUS_cor <-Heatmap(corr,
       name = "Correlation of Items",
       col= heatc,
       cluster_rows = TRUE,   # turn off clustering
       cluster_columns = TRUE,
       column_title = "Heatmap - Item Correlation")
plot(BUS_cor)

```
```{r}

```

##Confirmatory factor analysis

In the next section, the groups that have been identified in the heatmap are used to perform a confirmatory factor analysis (cfa).
For this purpose:
  1) the required packages are loaded
  2)load dataset
  3) the data is prepared
  4) checking assumptions of normality, additive, linearity and homogeneity
  5) create and run factor models
```{r load required packages}
```

```
if (!require("lavaan")) {
  install.packages("lavaan", dependencies = TRUE)
  library(lavaan)
}
if (!require("semPlot")) {
  install.packages("semPlot", dependencies = TRUE)
  library(semPlot)
}
if (!require("CTT")) {
  install.packages("CTT", dependencies = TRUE)
  library(CTT)
}
if (!require("haven")) {
  install.packages("haven", dependencies = TRUE)
  library(haven)
}
if (!require("psych")) {
  install.packages("psych", dependencies = TRUE)
  library(psych)
}
```

```{r load dataset}
CA_CFA_sav<-read_sav("C:\\Users\\Alison
Wang\\Desktop\\Bachelor\\analysis_scripts\\files\\CA_CFA_sav.sav")

```

```{r formating}
options("scipen"=100, "digits"=1) #disabeling use of exponential notation
```

```{r check assumptions}
#new column scores
CA_CFA_sav$scores <-rowSums(CA_CFA_sav[,c(12:26)])

CTT <- cfa_data %>%
  as.matrix()%>%
  CTT::itemAnalysis()

CTT$alpha
CTT$itemReport

#histogram of total scores
CA_CFA_sav %>%
  ggplot(aes(x = scores)) +
  geom_histogram(aes(y= ..density..), bins=30) +
  stat_function(
```

```
    fun = dnorm,
    args = list(mean = mean(CA_CFA_sav$scores),
                sd = sd(CA_CFA_sav$scores)))
hist(CA_CFA_sav$scores)

#skewness and kurtosis of the total scores
psych::skew(CA_CFA_sav$scores)
psych::kurtosi(CA_CFA_sav$scores, type = 3)

#Kaiser-Meyer-Olkin measure
cfa_data %>%
  KMO()


#Barlett's sphericity test
cfa_data %>%
  cortest.bartlett()

#multivarate outliers?
mahal = mahalanobis(cfa_data, colMeans(cfa_data),
                    cov(cfa_data, use = "pairwise.complete"))

summary(mahal)

#density plot of total scores/normality
CA_CFA_sav %>%
  ggplot(df, mapping= aes(x=scores)) +
  geom_density() +
  stat_function(
    fun = dnorm,
    args = list(mean = mean(CA_CFA_sav$scores),
                sd = sd(CA_CFA_sav$scores)))

#shapiro test
shapiro.test(CA_CFA_sav$scores)


```


```{r factor analysis}
##  defining six factor model - first order model
model_six <-
'ACC =~ Item_1+Item_2
EXP =~ Item_3+Item_5
UND =~ Item_4+Item_6
INF =~ Item_7+ Item_8 + Item_9
PRIV =~ Item_10
TIME =~ Item_11'
```

```
fit_six<-cfa(model_six, data=cfa_data, estimator="MLR")
factorload_six <- inspect(fit_six,what="std")$lambda  #save matrix with factor loadings

T.boot <- bootstrapLavaan(fit_six, R=10, type="bollen.stine",
                          FUN=fitMeasures, fit.measures="chisq")
#graphic
semPaths(fit_six,whatLabels="std",layout="tree")

## define hierachical second order model
model_hier <-'ACC =~ Item_1+Item_2
EXP =~ Item_3+Item_5
UND =~ Item_4+Item_6
INF =~ Item_7+Item_8 + Item_9
PRIV =~ Item_10
TIME =~ Item_11
QUAL=~ EXP+UND+INF'


fit_hier<-cfa(model_hier, data=cfa_data, estimator="MLR")
factorload_hier <-inspect(fit_hier,what="std")$lambda
#summary(fit_hier)

#graphic
semPaths(fit_hier,whatLabels="std",layout="tree")


##four factor model
model_four <-
'FRI =~ Item_1+ Item_2 + Item_11
EXP =~ Item_3 +Item_5 + Item_4 + Item_6
INF =~ Item_7+ Item_8 + Item_9
PRIV =~ Item_10'


fit_four<-cfa(model_four, data=cfa_data, estimator="MLR")
factorload_four <-inspect(fit_four,what="std")$lambda

##five factor model
model_five <-
'FRI =~ Item_1+ Item_2
EXP =~ Item_3 +Item_5 + Item_4 + Item_6
INF =~ Item_7+ Item_8 + Item_9
PRIV =~ Item_10
TIME =~  Item_11'

fit_five<-cfa(model_five, data=cfa_data, estimator="MLR")
factorload_five <-inspect(fit_five,what="std")$lambda

##three factor model
model_three <-
```

```
  'FRI =~ Item_1 + Item_2 + Item_11
QUAL =~ Item_3 + Item_4 + Item_5 + Item_6 + Item_7 + Item_8 + Item_9
PRIV =~ Item_10'



fit_three<-cfa(model_three, data=cfa_data, estimator="MLR")
factorload_three <-inspect(fit_three,what="std")$lambda

##Borsci 4 factor model
model_borscifour <-
  'one =~ Item_1 + Item_2
two =~ Item_3 +Item_4 +Item_5+ Item_6+Item_7+Item_8+Item_9
thr =~ Item_10
fou =~ Item_11'

fit_borscifour<-cfa(model_borscifour, data=cfa_data, estimator="MLR")
factorload_borscifour <-inspect(fit_borscifour,what="std")$lambda

##Borsci 5 factor model
model_borscifive <-
  'one =~ Item_1 + Item_2
sec =~ Item_3 +Item_4 +Item_5
two =~ Item_6+Item_7+Item_8+Item_9
fou =~ Item_10
fiv =~ Item_11'



fit_borscifive<-cfa(model_borscifive, data=cfa_data, estimator="MLR")
factorload_borscifive <-inspect(fit_borscifive,what="std")$lambda

fitMeasures_all <-sapply(list(fit_six,fit_hier,fit_four, fit_five, fit_three, fit_borscifour,
fit_borscifive), fitMeasures, c("chisq", "df","pvalue", "cfi",
"rmsea","rmsea.ci.lower","rmsea.ci.upper", "srmr", "aic", "bic", "bic2" ))
#create matrix with all fitmeasurements
colnames(fitMeasures_all) <- c('model six', 'model hier','model four', 'model five', 'model
three', 'model borscifour', 'model borscifive')       #name the columns of the matrix
print(fitMeasures_all)
```
```{r}
Group_names <- read_excel("C:/Users/Alison
Wang/Desktop/Bachelor/data/Group_names.xlsx")
table(Group_names$`Group names`)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot