

**Validating the Bot Usability Scale for Satisfaction with Chatbots: Factor
Structure, Convergent Validity and Relation with Workload**

Mustafa A. Taha

Faculty of Behavioural, Management and Social Sciences (BMS), University of

Twente

Human Factors & Engineering Psychology

1st Supervisor: Simone Borsci

2nd Supervisor: Jule Landwehr

July 6, 2022

Abstract

So far, no standardised scale to measure chatbot satisfaction is in widespread use, hindering research in this field. For this reason, the Bot Usability Scale (BUS) was developed. The aim of this study was to replicate and extend the validation of the BUS. Therefore, the main objectives were to confirm the BUS' previously established five-factor structure, high internal consistency, and strong positive correlation with the UMUX-Lite. A further objective was to test whether the BUS correlates negatively with workload, since workload and satisfaction typically have an inverse relationship.

For this, 58 participants solved tasks for five chatbots and rated each chatbot on the BUS, UMUX-Lite and NASA-TLX in an online study. In total, 388 BUS-questionnaires were collected and analysed using confirmatory factor analysis and Cronbach's alpha. For 297 out of all 388 BUS-questionnaires, corresponding UMUX-Lite and NASA-TLX questionnaires were also collected. Their relationship with the BUS was analysed using Kendall's Tau and linear mixed models. As expected, the five-factor structure was confirmed, the BUS had a high internal consistency, a strong correlation with the UMUX-Lite, and a negative relationship with workload.

Overall, these findings support the use of the BUS as a standardized measure for chatbot satisfaction, enabling replicable research on this topic, comparison of chatbots, and the establishment of benchmarks for desirable satisfaction levels. However, the factor analysis revealed a variance of 0 [-0.18, 019] for item 1, suggesting a potential Heywood case which may require modification or removal. Future research should aim to replicate these findings with larger, simple random samples, with chatbots with overall low satisfaction rates and high workload scores and using multiple tasks per chatbot. Moreover, future studies should research whether item 1 presents a Heywood case, if so, why it occurs and whether modification or removal of item 1 is appropriate. Finally, researchers should investigate the relationships between the BUS and variables other than workload.

Table of Contents

Introduction	4
Method	10
Participants	10
Materials	10
Procedure	11
Data Analysis	12
Results	14
Discussion	20
Conclusion	24
References	25
Appendix A. Demographic Questions	31
Appendix B. Chatbots and Task Scenarios	32
Appendix C. BUS-Scale with Instructions	35
Appendix D. UMUX-Lite with Instructions	36
Appendix E. Adapted NASA-TLX with Instructions	37
Appendix F. Study Information and Informed Consent	39
Appendix G. General Instructions	40
Appendix H. R Script for Data Analysis	41

Introduction

In times of ongoing digitalization, an increased number of organizations have used chatbots to move their services online. For instance, chatbots are used to offer automated customer service (Nicolescu & Tudorache, 2022), support students with educational materials and personalized help, or assist patients by providing health information and reminders for treatments (Adamopoulou & Moussiades, 2020). In short, chatbots are used to perform a variety of tasks in online services.

A clear definition of chatbots is needed to research them. A chatbot is defined as software that chats or interacts with users using natural language (Brandtzaeg & Følstad, 2018; Klopfenstein et al., 2017). When first invented in the 60s, the chatbot simply detected keywords in the users' input and transformed them based on a specific given rule (Weizenbaum, 1966). Nowadays, chatbots may use AI to recognize themes and goals to allow for a more flexible conversation but less complex methods are still in use (Gupta et al., 2020). Visually, chatbots often take the appearance of messenger apps to give the impression of a conversation with a real person (Klopfenstein et al., 2017). In general, chatbots can have a broader focus and can handle various topics. In customer service, however, chatbots are typically more narrow and only able to maintain conversations related to specific tasks (Grudin & Jacques, 2019). To summarize, chatbots are conversational software which differ in the technology they use and the range of conversational topics they can handle.

Research on chatbots is important because they are increasingly used in everyday life. To illustrate, they are frequently used in multiple industries (Behera et al., 2021), such as tourism (Lasek & Jessa, 2013), real estate (Quan et al., 2019), health care (Hwang et al., 2020), e-commerce and education (Caldarini et al., 2022). Moreover, customers seem to have a positive attitude towards them. In fact, Tran et al. (2021) report that in some instances, customers feel more positive about their interactions with chatbots than with human agents, increasing their relevance in business and thus, research. Further, they are convenient for both businesses and the end-user as, they provide a cheap, easy-to-use tool to obtain or offer service- and product-related information in real-time (Behera et al., 2021). In short, research on chatbots is important as they play a major and growing role in customer service for many industries, which use them as a cheap and convenient messaging tool for customers.

To enable replication and comparison of research on chatbots, standardized measurements for aspects of the chatbot's interaction quality are needed. In this context, a key factor for assessing the quality of chatbots is the user's satisfaction, which is part of a chatbot's usability. The International Organization for Standardization (ISO) 9241-11 defines

usability as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO 9241-11, 2018). ISO 9241-11 (2018) further defines satisfaction as “positive attitudes, emotions and/or comfort resulting from use of a system, product or service”. This means that satisfaction encompasses subjective factors instead of the mere ability to achieve a certain goal. For chatbots, this means the user should not only arrive at their desired output but also have a satisfying experience during the interaction. In brief, satisfaction is an aspect of a chatbot’s interaction quality and thus, research on chatbots would benefit from a standardized measure for satisfaction with chatbots.

There are several reasons why user satisfaction is important for chatbots in customer service and should not be neglected in research. For one, previous studies confirm that factors beyond effectiveness and efficiency influence users’ satisfaction with chatbots, making it important to look at as a separate factor. For example, emotional strategies such as reflecting the user’s emotion, expressing emulated emotions as a chatbot, and apologizing seem to have a positive effect on user satisfaction (Liu et al., 2020). Moreover, message interactivity, such as showcasing the memory of previous responses, and anthropomorphic visual cues, such as an icon of a human face, appear to affect the perceived social presence and humanness of the chatbot, further contributing to positive attitudes (Go & Sundar, 2019). Lastly, mental workload, which describes the degree of effort a user has to exert to complete a task (Hart & Wickens, 1990; Kantowitz, 1987), was found to have a negative relationship with chatbot satisfaction (Nguyen et al., 2022). Overall, one reason to study satisfaction as an independent concept is that there are various factors other than effectiveness and efficiency that influence a user’s experience with a chatbot.

Secondly, user satisfaction with chatbots affects the overall brand-customer-relationship, giving it relevance for a company’s brand image. For instance, the perceived communication quality and entertainment value of a chatbot can improve the customer-brand-relationship (Cheng & Jiang, 2021). Similarly, Kull et al. (2021) found that whether a chatbot initiates a conversation with a warm or competent message influences the consumers’ brand-relation and engagement. As shown by Cheng and Jiang (2021), the customer-brand-relationship affects purchasing decisions a customer makes. This means that user satisfaction with chatbots could be intended as a driver of purchasing decisions since it affects the brand-image and relationship. At last, satisfaction influences the user’s intention to continue using the chatbot (Ashfaq et al., 2020; Nguyen et al., 2021). So, satisfaction in the interaction with chatbots is needed if companies want to motivate customers to keep using their services. To

sum up, satisfaction is an important factor for chatbots in customer service because it influences the user's intention for future usage and affects the overall brand-relationship, driving purchasing decisions.

To facilitate the measurement of chatbot satisfaction, the first standardised scale for this construct was recently proposed. Borsci et al. (2021b) developed the Bot Usability Scale (BUS) (Table 1), which consists of 11 items on a five-point Likert scale and has five factors. Namely, the factors are (a) perceived accessibility to chatbot functions, (b) perceived quality of chatbot functions, (c) perceived quality of conversation and information provided, (d) perceived privacy and security and (e) time response (Borsci et al., 2021b). Briefly, the BUS measures chatbot satisfaction using 11 items across 5 factors.

Table 1

Factors and Items of the Bot Usability Scale (BUS)

Factor	Item
1: Perceived accessibility to chatbot functions	1: The chatbot function was easily detectable. 2: It was easy to find the chatbot.
2: Perceived quality of chatbot functions	3: Communicating with the chatbot was clear. 4: The chatbot was able to keep track of context. 5: The chatbot's responses were easy to understand.
3: Perceived quality of conversation and information provided	6: I find that the chatbot understands what I want and helps me achieve my goal. 7: The chatbot gives me the appropriate amount of information. 8: The chatbot only gives me the information I need. 9: I feel like the chatbot's responses were accurate.
4: Perceived privacy and security	10: I believe the chatbot informs me of any possible privacy issues.
5: Time response	11: My waiting time for a response from the chatbot was short.

Note. Retrieved from “Confirmatory Factorial Analysis of the Chatbot Usability Scale: A Multilanguage Validation.” by Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A. & van der Velde, F., 2021. [Manuscript submitted for publication].

Currently, the scale is still within the process of further validation. To begin its development, Borsci et al. (2021b) conducted a systematic review and online survey to identify relevant aspects of chatbot satisfaction and then used focus groups to create an initial set of items. Next, a pilot study was conducted to perform psychometric evaluations (Borsci et al., 2021b). In detail, an exploratory factor analysis revealed a five-factor structure and after dropping some less reliable items, the scale had an internal consistency of $\alpha = .87$ (Borsci et al., 2021b). Furthermore, the scale correlated with the previously established satisfaction scale UMUX-LITE, indicating external validity (Borsci et al., 2021b). However, some factors only correlated mildly with the UMUX-LITE, indicating that the new scale covers satisfaction more broadly (Borsci et al., 2021b). Afterwards, other studies further confirmed external validity with the UMUX-LITE (Lopez, 2021; Waldmann, 2021) and the Speech User Interface Service Quality (SUISQ-R) (van den Bos, 2021). Again, they found that the new questionnaire seemed to include more factors, as would be expected from a scale that also covers the conversational aspects of chatbots. In contrast to Borsci et al. (2021b), Lopez (2021) and van den Bos (2021) found only four factors, whereas Waldmann (2021) found six factors. However, Borsci et al. (2021a) recently found that the current version of the BUS consists of five factors. In conclusion, the scale appears to have a five-factor structure, a good level of reliability and it is correlated with other satisfaction questionnaires while covering additional and relevant aspects associated with the quality of interaction with chatbots.

The BUS was developed to provide researchers with a standardized scale to measure chatbot satisfaction. Before its development, scales for Human-Computer-Interaction satisfaction were used for chatbots (Borsci et al., 2021b). However, conversational and interactional aspects of chatbots are not covered in these scales, so they were supplemented with qualitative methods (Borsci et al., 2021b). As a result, it was difficult to compare and replicate studies, compare chatbots and establish benchmarks for desirable satisfaction levels for different uses (Borsci et al., 2021b). Thus, the BUS facilitates standardized research on chatbot satisfaction.

Still, further testing is needed before using the BUS as a standard tool in research. Accordingly, this study will investigate the BUS' factor structure, reliability, construct validity and relation with workload to improve the BUS' validation. For one, a scale

validation should be conducted using multiple samples to assess the generalisability of the scale's properties (Rauvola et al., 2020). Therefore, additional studies should replicate key findings of the BUS' characteristics. For instance, the BUS' factor structure and reliability are important attributes, as these analyses inform researchers about the dimensionality and internal consistency of a scale, which indicate a scale's quality (Lamm et al., 2020; Zhou, 2019). Thus, this study will aim to replicate the BUS' five-factor structure and high reliability to determine the generalisability of these characteristics.

In addition, the BUS' construct validity will be assessed. Construct validity, which is the degree to which a scale measures the construct that it is intended to measure, is crucial to a scale's validity (Cronbach & Meehl, 1955). An approach for assessing this is called convergent validity testing, which is used to describe the degree to which a scale correlates with another measure of a similar construct (Rauvola et al., 2020). In the case of the BUS, this means that its correlations with other satisfaction scales are a critical indicator of validity. One such measure is the UMUX-Lite (Lewis et al., 2013), which has two items on a seven-point Likert scale and was developed as a short version of the UMUX, another satisfaction questionnaire. It only includes UMUX-items with a positive tone, which ask about a user's positive rather than negative attitudes towards a system (Lewis et al., 2013). Furthermore, the UMUX-Lite has good psychometric properties. Specifically, the UMUX-Lite has high reliability estimates, with Cronbach's alpha ranging from .7 (Borsci et al., 2020) to .83 (Lewis et al., 2013), and is strongly correlated with other satisfaction measures, such as the System Usability Scale and the UMUX (Borsci et al., 2015). This robustness makes the UMUX-Lite a suitable measure to establish the BUS' convergent validity. Previous studies on the BUS have found that its relationship with the UMUX-Lite matches expectations, as they have a moderate overall correlation but some BUS factors correlate more strongly, suggesting that the BUS covers more facets than the UMUX-Lite, as intended (Borsci et al., 2021b; Lopez, 2021; Waldmann, 2021). Due to the importance of this type of validity and the suitability of the UMUX-Lite as a comparative measure, this study will aim to replicate these findings. To sum up, the BUS' construct validity will be investigated by measuring the correlation between the BUS and the UMUX-Lite.

Finally, it is good practice to study a new scale's relationships to other variables related to the scale's construct for more complete validation. Rauvola et al. (2020) refer to this process as the establishment of a nomological network, which represents how the scale is related to various constructs which should relate to the construct that the scale is intended to measure. Following this, the BUS' relationships to variables which usually relate to

satisfaction must be studied. For example, workload has a negative relationship with satisfaction in many settings. According to the data of Karczewska et al. (2021), users tend to be less satisfied with mobile apps that are demanding in terms of workload, compared to when they have to interact with systems that require a low workload for the interaction. Similarly, Schmutz et al. (2009) and Mirhoseini et al. (2021) observed a negative correlation between user satisfaction and workload in online shopping. Moreover, this negative relationship between satisfaction and workload has recently been confirmed for chatbots (Nguyen et al., 2022). Following this, it is expected that the BUS should also have a negative relationship with measures of workload. Thus, this study will assess the relationship between the BUS and workload as part of the establishment of a nomological network.

To test the predicted relationship between the BUS and workload, a suitable measure of workload is necessary. In this context, the NASA-Task Load Index (NASA-TLX) can be used. This scale measures workload in six dimensions, including mental demand, physical demand, temporal demand, performance, effort, and frustration (National Aeronautics and Space Administration (NASA), n.d.). First, participants are shown a definition of each dimension (NASA, n.d.). Then, participants rate each dimension by marking a score on a line with “0” on the left and “100” on the right (NASA, n.d.). Further, the relative weight of each dimension is determined by the participant through pairwise comparisons. This means that the participant chooses which of the two dimensions had a stronger influence on the workload of the system or task, repeating this procedure for all possible pairs of dimensions, thus ranking the dimensions by importance (NASA, n.d.). Finally, the ratings are combined into an overall score for workload (NASA, n.d.). The NASA-TLX is an appropriate measure to test the relationship between the BUS and workload for multiple reasons. For one, it is a commonly used measure (Ruiz-Rabelo et al., 2015) that was found to be valid in many different contexts, such as monitoring health patients (Said et al., 2020), learning difficult surgical methods (Ruiz-Rabelo et al., 2015), and measuring workload in older adults (Devos et al., 2020). Secondly, Xiao et al. (2005) report that the NASA-TLX has high reliability and desirable structure validity. Lastly, the NASA-TLX covers multiple facets of workload because it includes six different dimensions. Thus, it allows for a more complete measurement of workload, which is commonly considered a complex construct (Hart & Wickens, 1990; Kantowitz, 1987). Hence, this study will use the NASA-TLX as a measure of workload because it is commonly used, validated, reliable, and because it encompasses multiple facets of workload.

In summary, this study aims to investigate multiple aspects associated with the interaction with chatbots.

To further confirm the psychometric properties of the BUS emerged from previous studies, the present work will (a) attempt to confirm the previously established five-factor structure of the BUS (R1); (b) assess the reliability of the BUS scale expecting at least a Cronbach's alpha above 0.85 for the 11 items (R2), and (c) confirm the strong correlation between the BUS and UMUX-Lite identified in previous studies (R3). Moreover, this work will investigate the relationship between the BUS and workload, as assessed by the NASA-TLX, with the expectation that there will be a negative relationship between these scales (R4).

Method

A within-subjects design was adopted asking participants to assess their satisfaction (BUS, and UMUX-Lite) and their perceived workload (NASA-TLX) on a set of chatbots.

Participants

Participants were recruited for the study if they were fluent in English and either had no visual impairment or had corrected vision. The participants were recruited in multiple ways: Some participants were students at the University of Twente and received credit points for their participation, which they needed to complete their study programme. In addition, acquaintances of the researcher were asked to participate and the study was promoted on social media and survey sharing websites.

109 people participated in the study, of which 58 (53.21%) completed the questions for all 5 chatbots and 51 (46.79%) completed the questions for between one and four chatbots. In total, 388 observations (e.g. complete questionnaires) were included in the confirmatory factor analysis and computation of Cronbach's alpha. For 46 participants, data on the UMUX-Lite and NASA-TLX was not collected or not collected correctly, as this data was taken from a different, related study which measured the UMUX-Lite differently and did not use the NASA-TLX. Therefore, only data of 63 people with 297 total observations were included in the correlational analyses and linear mixed model analyses.

Demographic data on the participants' age, sex, English proficiency and previous experience with chatbots was collected. The participants' ages ranged between 15 and 54 years ($M = 24.58$; $SD = 8.06$). Of all participants, 44.04% were male ($n = 48$), whereas 55.05% were female ($N = 60$) and 0.01% ($N = 1$) did not disclose their sex. English proficiency varied, with five people indicating basic (A1 - A2) proficiency, 25 people

reporting intermediate (B1 – B2+) proficiency, 61 people indicating advanced (C1 – C2) proficiency and 18 native speakers. Most participants had high chatbot experience, which had a mean of 77.8% ($SD = 20.8\%$).

Materials

The data was collected in Qualtrics, an online survey software that enables researchers to display items in various formats and gather data.

To collect demographic information, participants were asked about their age in years, their English proficiency, sex assigned at birth, and prior experience with chatbots (See Appendix A). English proficiency was categorized as basic (A1 - A2), intermediate (B1 – B2+), advanced (C1 – C2) or native speaker, whereas chatbot experience was assessed using three items on five-point Likert scales.

Seven customer service chatbots were included in the study. Each chatbot belonged to a different company or institution (see Appendix B). For each chatbot, a specific task was designed, which was solvable and of similar difficulty as the tasks for other chatbots (see Appendix). As an example, the customer service chatbot on Samsung’s website was included with the following task: “You live in the USA and have ordered a TV by Samsung. However, the image is flickering, so you want to chat with Samsung’s support. Once you have found their chatbot, use it to schedule a repair for your flickering TV. Your task is done once the chatbot gives you the option to schedule a repair, so you can stop without really making an appointment.”

Satisfaction with the chatbots was measured using the BUS (see Appendix C) and the UMUX-Lite (see Appendix D). To measure workload, the NASA-TLX (See Appendix E) was used. Its instructions were adapted to emphasize that participants should rate the workload of the chatbot, not the task. To reduce the time needed to fill in the NASA-TLX, its instructions were shortened and the weighing of its six dimensions was omitted. Instead, all dimensions contributed equally to the overall workload score.

Procedure

The researcher sent a link to the participants, which they could use to access the study in Qualtrics. Here, they were first provided with a page presenting information about the study, including its aims, activities, expected duration, data collection and management, expected risks, right to withdraw and contact information of the researchers (see Appendix F). At the bottom of this page, participants indicated whether they agreed to take part in the

study. If they agreed, they were asked to provide demographic information. Following this, they were randomly allocated to assess 5 chatbots out of the 7 chatbots included in the study and given general instructions on how to perform the tasks (see Appendix G). In detail, they were asked to read the task scenario, then find the chatbot using the provided weblink, and fulfil the task before proceeding to the questionnaires. In case some participants could not complete all tasks, they were instructed to proceed with the questionnaire if they could not finish a task within 10 minutes. In addition, they were instructed not to offer personal information to the chatbots, use them to schedule appointments, or contact human employees. After the participants had read this information, the link to the first chatbot with a specific task was shown. Once the participants felt they completed the task, they proceeded to fill in the BUS-Scale, UMUX-Lite and then the NASA-TLX. Then, the next chatbot and task were presented and the procedure was repeated until the participant had filled in the questionnaires for all 5 chatbots. Finally, the participants were thanked for their participation and reminded of the researchers' contact information in Qualtrics.

Data Analysis

The data was exported from Qualtrics XM (Qualtrics, 2005) as a CSV-file with numeric values, which was imported into R (R Core Team, 2022), which was used for the entire data analysis (see Appendix H). Here, unnecessary variables and observations were removed, including metadata, data on chatbots with incomplete BUS, UMUX-Lite or NASA-TLX questionnaires and data from participants who did not give consent. Furthermore, the data was reorganized into long format, so there was one observation for each pair of participants and chatbots. Finally, total scores for the BUS and its factors, the UMUX-Lite and the NASA-TLX were computed. For this, the mean of all items belonging to a scale was calculated and subsequently divided by the highest possible score. This way, all scores were transformed to a scale up to 1, which made comparisons between the scores easier to interpret.

For the data analysis, we first looked at confirming the psychometrics properties of the BUS (R1 to R3) and then at the relationship between the BUS and the NASA-TLX (R4)

In order to investigate the factorial structure of the BUS, a confirmatory factor analysis using the R package "lavaan" (Rosseel, 2012) was performed. Prior to the analysis, the normality of each BUS item was assessed using the Shapiro-Wilk test, whereby items with $p > .05$ were considered normally distributed. The factorial model was specified to match the findings of Borsci et al. (2021a) (Table 1) and estimated using robust maximum likelihood

(MLR). MLR is an alternative to maximum likelihood (ML), the most common method for estimating factorial models (Li, 2016; Maydeu-Olivares, 2017). In contrast to ML, MLR can be used for non-normally distributed data (Li, 2016; Maydeu-Olivares, 2017; Pavlov et al., 2020), making it appropriate for more data sets. Model fit was evaluated using the following benchmarks: Chi-square should ideally have a p-value above .05, but at least above .001. In addition, the comparative fit index (CFI) should be above 0.9, the root mean square error of approximation (RMSEA) should ideally be below 0.05 but at most 0.08, and the standardized root mean square residual (SRMR) should be equal to or smaller than 0.05. Further, standardized factor loadings and the variance of each item were inspected to see how strongly individual items correlated with its factor and how much of the variance was explained by the factor model. Finally, the factor model was represented graphically using the R package “semPlot” (Epskamp et al., 2022).

The Cronbach’s alpha for the overall BUS and for each factor that contained more than one item was computed using the R package “ltm” (Rizopoulos, 2006).

Correlations of the BUS and its factors with the UMUX-Lite and NASA-TLX were calculated using Kendall’s Tau with a significance level of $p \leq .05$. This measure was chosen because it can be used for non-normally distributed variables. Shapiro-Wilk tests for the overall BUS, UMUX-Lite, and NASA-TLX scores were used to assess whether these variables were normally distributed. Further, these correlations were plotted using the R package “ggplot2” (Wickham, 2016).

Moreover, to investigate the relationship between the scales, linear mixed models were performed using the R package lme4 (Bates et al., 2015) and p-values were obtained with the “lmerTest” package (Kuznetsova et al., 2017). This analysis was conducted to study whether correlations between the BUS and other variables could be observed when accounting for the repeated measures of each participant and chatbot. Each model used the BUS score or one of its factors as the independent variable, while the UMUX-Lite or NASA-TLX was the dependent variable. Additionally, random effects for the participants or chatbots were included in the model if they led to clustering in the dependent variable. To identify clustering, boxplots were used. Then, the selected linear mixed model was estimated and its assumptions were tested. For this, scatterplots of the independent and dependent variables were created and the distribution of the residuals was plotted. If the assumptions were met, the slope coefficient of the independent variable was used as an indicator of the relationship between the variables, using a significance level of $p \leq .05$.

Results

Descriptive statistics of the main outcome variables are shown in Table 2. Overall, the BUS and UMUX-Lite scores were skewed towards the upper end of the scale, whereas the NASA-TLX scores were skewed towards the lower end.

Table 2

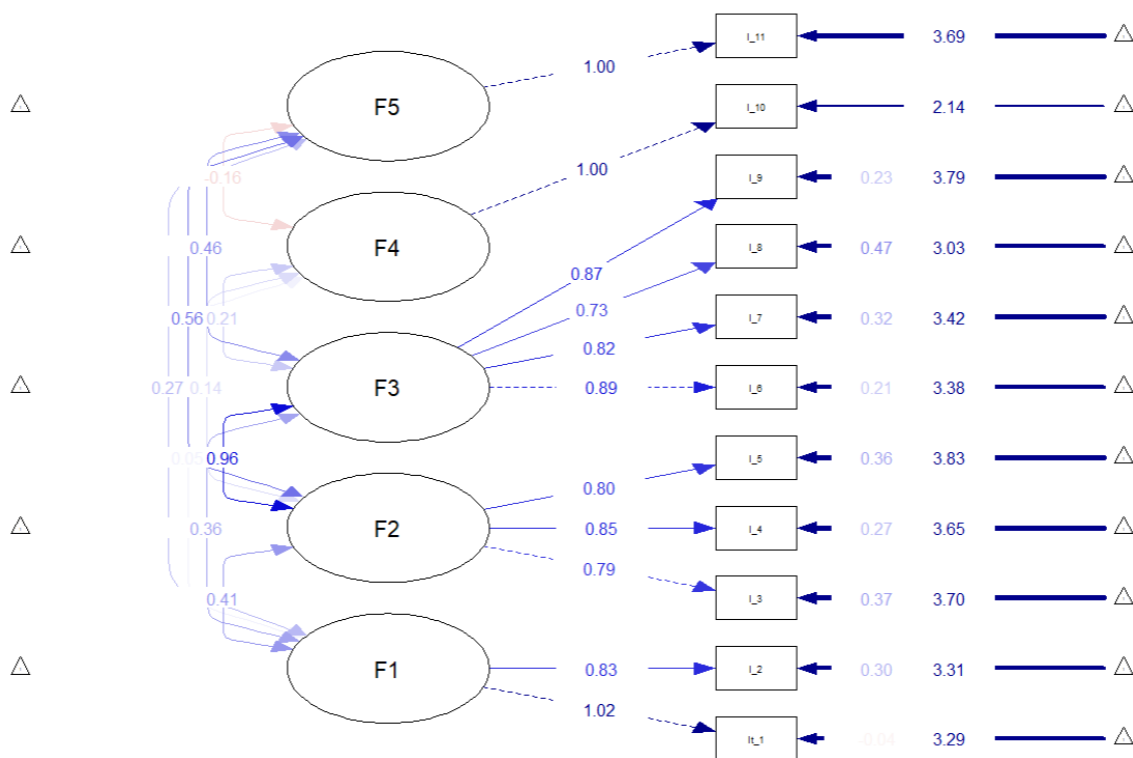
Quartiles, Medians, Means and Standard Deviations of the BUS, UMUX-Lite, and NASA-TLX Scores

Variable	1 st Quartile	Median	3 rd Quartile	Mean	Standard Deviation
BUS	69%	78%	87%	76%	15%
Factor 1	70%	80%	100%	79%	23%
Factor 2	73%	80%	93%	80%	18%
Factor 3	65%	80%	95%	77%	20%
Factor 4	40%	60%	60%	53%	25%
Factor 5	80%	80%	100%	82%	22%
UMUX-Lite	64%	86%	93%	79%	22%
NASA-TLX	10%	23%	46%	28%	21%

The Shapiro-Wilk test suggested that none of the BUS-items was normally distributed, with $p < .001$ for each item. Thus, MLR was used to estimate the factorial model (Figure 1), since it can be used for non-normally distributed variables (Li, 2016; Maydeu-Olivares, 2017; Pavlov et al., 2020).

Figure 1

The Five Factor Model with Factor Loadings, Covariance between Factors, and Item Variance



Most of the fit measure outcomes indicated good model fit (Table 3). In detail, the CFI (0.959) and SRMR (0.03) met the pre-specified benchmarks. The RMSEA (0.074) is in an acceptable range for good model fit but would optimally be lower. On the other hand, Chi-square ($\chi^2 = 112.206$, $p < .001$) is significant, indicating poor model fit. However, Chi-square is considered a less reliable fit index because it is strongly affected by sample size and may over reject factor models (Alavi et al., 2020; Hutchinson & Olmos, 2009; Li, 2016). Therefore, to answer our research question (R1), it appears the factor model is confirmed.

Table 3*Fit Measure Values and Benchmarks, and Assessment of Model Fit*

Fit Measure	Value	Benchmark	Assessment of Model Fit
Chi-square	112.206, $p < .001$	$p > .001$, ideally $p > .05$	Poor
Comparative Fit Index (CFI)	0.959	CFI > 0.9	Good
Root Mean Square Error (RMSEA)	0.074	RMSEA < 0.08, ideally RMSEA < 0.05	Acceptable
Standardized Root Mean Square Residual (SRMR)	0.03	SRMR \leq 0.05	Good

The standardized factor loadings and variances (See Table 4) provide insight into how the items are related to the factors. Based on the standardized factor loadings, there were strong relations between most of the items and the factor model (see Table 4). Items 1, 10, and 11 had a standardized factor loading of 1 or higher, indicating that they contributed very strongly to their factor. For items 10 and 11, this was inevitable, as they constitute the single-item factors 4 and 5. Of the remaining items, most have standardized factor loadings of 0.8 or higher, which suggests that they also correlate strongly with their factors. Only items 3 and 8 have somewhat weaker relationships with their factors, with standardized factor loadings of 0.76 and 0.72, respectively. The variances of the items were all positive, although the variance of item 1 approximated 0 with a value of 0.002, CI [-0.18, 0.19] (see Table 4). This suggests a possible Heywood case for item 1, meaning that the estimated variance may be negative (Kolenikov & Bollen, 2012). Overall, the standardized factor loadings and variances indicate desirable relations between the items and the factor structure except for item 1, which may present a Heywood case.

Table 4

Standardized Factor Loadings, Standardized Variance, and Confidence Intervals of Unstandardized Variance of BUS-Items

Factor	Item	Std. Factor Loading	Std. Variance	Variance – Lower Bound	Variance – Upper Bound
1: Perceived accessibility to chatbot functions	1	1	0.002	-0.18	0.19
	2	0.86	0.27	0.21	0.55
2: Perceived quality of chatbot functions	3	0.76	0.42	0.37	0.56
	4	0.8	0.36	0.31	0.53
	5	0.81	0.35	0.28	0.46
3: Perceived quality of conversation and information provided	6	0.86	0.27	0.26	0.44
	7	0.81	0.35	0.36	0.54
	8	0.72	0.49	0.57	0.88
	9	0.87	0.24	0.19	0.31
4: Perceived privacy and security	10	1	0	0	0
5: Time response	11	1	0	0	0

The reliability of the overall BUS and its multi-item factors was high. As a whole, the BUS had a reliability of $\alpha = 0.88$ with 11 items, which meets expectations. Factor 1 had the highest reliability, with $\alpha = 0.92$ and two items, followed by factor 3 with four items ($\alpha = 0.89$) and factor 2 with three items ($\alpha = 0.84$). Therefore, to answer our research question (R2) the BUS and its factors appear to have good internal consistency.

Kendall's Tau indicated strong correlations between the BUS and the UMUX-Lite. Based on the Shapiro-Wilk test, the total scores of the BUS, its factors, and the UMUX-Lite were not normally distributed ($p < .001$). Therefore, the use of a non-parametric correlation analysis such as Kendall's Tau was appropriate to investigate the correlations between the scales in line with our research question (R3). This analysis revealed significant correlations between the UMUX-Lite and the BUS scores except for factor 4 (see Table 5). In short, the research question (R3) can be answered by stating that except for factor 4, the BUS and its factors correlate with the UMUX-Lite based on Kendall's Tau.

Table 5*Correlation between the BUS-Subscales and the UMUX-Lite*

BUS- Subscale	Kendall's Tau	p-value
BUS	0.66	< .001
Factor 1	0.35	< .001
Factor 2	0.68	< .001
Factor 3	0.6	< .011
Factor 4	0.03	0.58
Factor 5	0.46	< .001

The linear mixed model analysis led to similar findings as Kendall's Tau. The UMUX-Lite scores were clustered by both the participants and the chatbots. Therefore, linear mixed models with random effects for participants and the chatbots were estimated. The assumptions of the models were mostly met, although all models had a slight overrepresentation of negative residuals at high fitted scores for the UMUX-Lite and more relatively large residuals than expected for a perfectly normal distribution. For all BUS-sub-scales, the slope coefficients were significant (see Table 6). The overall BUS-score had the strongest correlation with a slope coefficient of 1.19, followed by factor 2 ($\beta = 0.95$) and factor 3 ($\beta = 0.84$). The remaining correlations were moderate, ranging from 0.19 to 0.44. Hence, to answer the research question (R3), a strong to moderate significant correlation between the BUS-sub-scales and the UMUX-Lite were found when including random effects for participants and the chatbots in the analysis.

Table 6*Outcomes of Linear Mixed Model Analysis between the BUS-Subscales and the UMUX-Lite*

BUS- Subscale	Slope Coefficient	t-statistic	p-value
BUS	1.19	23.74	< .001
Factor 1	0.36	6.52	< .001
Factor 2	0.95	22.45	< .001
Factor 3	0.84	19.25	< .001
Factor 4	0.19	3.68	< .001
Factor 5	0.44	8.11	< .001

Regarding the final research question (R4), Kendall's Tau revealed expected correlations of the BUS and its factors with workload, except for factor 4. The NASA-TLX responses did not follow a normal distribution based on the Shapiro-Wilk test, which returned a p-value below .001. Thus, a non-parametric test like Kendall's Tau was necessary for the correlational analysis. Kendall's Tau was significant for the relationship between all BUS-subscales and the NASA-TLX (Table 7). Most subscales had moderate negative correlations with the NASA-TLX with slope coefficients between -0.22 and -0.4, which is in line with expectations. In contrast, factor 4 "Perceived privacy and security" had a small, positive correlation ($\beta = 0.12$) with the NASA-TLX, which was not expected (R4). As such, it seems there is a negative relationship between the BUS or most of its factors and workload, whereas factor 4 and workload may be positively related.

Table 7

Correlation between the BUS-Subscales and the NASA-TLX

BUS- Subscale	Kendall's Tau	p-value
BUS	-0.36	< .001
Factor 1	-0.22	< .001
Factor 2	-0.4	< .001
Factor 3	-0.35	< .001
Factor 4	0.12	.005
Factor 5	-0.34	< .001

For the most part, the outcomes of the linear mixed model analysis matched the findings using Kendall's Tau. Boxplots of the NASA-TLX scores revealed that they differed between participants and chatbots. For this reason, linear mixed models with random effects for the participants and chatbots were estimated. All models largely complied with the model assumptions but had a somewhat increased number of residuals at the extremes of the residual distribution and slightly smaller residual variance at low fitted levels of the NASA-TLX. The slope coefficients of the BUS subscales were all negative and varied in size (Table 8). Except for factor 4, all slope coefficients were significant ($p < .001$). Overall, the research question (R4) can be answered by stating that a significant negative relationship was found between the BUS or its factors, except factor 4, and the workload measured by the NASA-TLX when accounting for clustering in the data.

Table 8*Outcomes of Linear Mixed Model Analysis between the BUS-Subscales and the NASA-TLX*

BUS- Subscale	Slope Coefficient	t-statistic	p-value
BUS	-0.62	-10.9	< .001
Factor 1	-0.16	-3.63	< .001
Factor 2	-0.49	-11.07	< .001
Factor 3	-0.39	-8.89	< .001
Factor 4	-0.05	-1.08	0.283
Factor 5	-0.28	-6.83	< .001

Discussion

The aim of this study was to continue the validation of the BUS by confirming previous findings on its psychometric properties and confirming its strong positive correlation with the UMUX-Lite. Further, the relationship between the BUS and workload was tested with the expectation of a negative correlation.

First, a confirmatory factor analysis was performed to replicate the five-factor structure found by Borsci et al. (2021a). While most fit indices confirmed this model, Chi-square rejected it. However, Li (2016) observed that Chi-square tends to over reject factor models when samples include fewer than 500 observations, which was the case in the present study. Therefore, Chi-square may have rejected the five-factor structure because of the number of observations rather than poor model fit. Other authors also report that Chi-square is affected by sample size and tends to over reject factor models (Alavi et al., 2020; Hutchinson & Olmos, 2009). Therefore, it is recommended to use multiple fit measures and not use Chi-square as the sole basis for rejecting a model (Alavi et al., 2020). Using these guidelines, the five-factor model can be confirmed, as all the other fit measures indicated a good model fit.

In the parameter estimates of the model, there was an unusual observation for item 1. This item had a variance of 0, indicating that it does not contribute to the scores of factor 1. Possibly, this may be because of its similarity with another item 2 “It was easy to find the chatbot”, which is close in meaning to item 1 “The chatbot function was easily detectable”. Hence, item 1 may not provide much or any additional information beyond item 2 and its exclusion may improve the factor model without information loss. Alternatively, item 1 may present a Heywood case, in which the estimated variance is negative (Kolenikov & Bollen, 2012). Heywood cases can occur for many reasons but are generally considered problematic because negative variances cannot truly occur in the population (Kolenikov & Bollen, 2012).

Thus, it should be investigated whether item 1's variance is truly negative and if so, what causes this, so the item can be removed or modified appropriately. In summary, the variance of item 1 indicates either a low or no contribution of this item to the factor or a Heywood case. Future research should investigate which of these alternatives is true and if it is a Heywood case, why it occurs and how it can be avoided.

Second, the internal consistency of the BUS was investigated. In line with expectations, the internal consistency of the scale was high, with $\alpha > 0.85$. This confirms previous research on the BUS, which found reliability estimates between 0.87 (Borsci et al., 2021b) and 0.97 (Lopez, 2021) for the overall BUS. In addition, the internal consistency of the individual BUS-factors was similarly high. Consequently, the factors are reliable enough to be used as subscales for specific aspects of chatbot satisfaction. For this, future studies are necessary to establish the validity of the factors on their own. In short, this study found that both the overall BUS and its factors have a sufficiently high internal consistency for the use as a satisfaction scale or subscale.

Third, this study aimed to replicate the positive correlation between the BUS and the UMUX-Lite to assess the BUS' convergent validity. Following expectations, a significant, strong, positive correlation between the BUS and the UMUX-Lite was found, with some factors correlating more strongly with the UMUX-Lite than others. The different strengths of the correlations can be explained by the content of the factors. For instance, factor 2 "Perceived quality of chatbot functions", which has items overlapping with the UMUX-Lite item "The chatbot is easy to use", had the strongest relationship with the UMUX-Lite. Similarly, the second strongest correlation was found with factor 3 "Perceived quality of conversation and information provided", which has items similar to the UMUX-Lite item "The chatbot's capabilities meet my requirements". Notably, Waldmann (2021) also found that these two factors of the BUS correlate most with the UMUX-Lite. Thus, it appears that factors 2 and 3 most strongly resemble the aspects of satisfaction measured in the UMUX-Lite. On the other hand, factors 1, 4, and 5 had weaker correlations with the UMUX-Lite, with factor 4 having a non-significant relationship according to Kendall's Tau but not the linear mixed model analysis. However, this does not mean that these factors do not validly measure aspects of chatbot satisfaction. Rather, "Perceived accessibility to chatbot functions", "Perceived privacy and security", and "Time response" may present facets of satisfaction which are relevant to chatbots but may not be related to many other systems. For this reason, they are included in the BUS but not the UMUX-Lite, explaining the weak or non-significant correlations. Thus, it appears that the BUS covers a wider definition of satisfaction than the

UMUX-Lite, as previous researchers have also found (Borsci et al., 2021b; Waldmann, 2021). In brief, this study confirmed the strong correlation of the overall BUS, factor 2 and factor 3 with the UMUX-Lite, whereas the other factors had weaker correlations with the UMUX-Lite. These findings indicate that the BUS is a valid measure of chatbot satisfaction, which is a broader construct than general satisfaction as measured by the UMUX-Lite.

Finally, this study tested the relationship between the BUS and workload. As expected, the BUS and factors 1, 2, 3, and 5 had a significant negative correlation with workload. Thus, the current findings match previous research on workload and satisfaction, which found a negative relationship between these variables for chatbots (Nguyen et al., 2022) and other online technologies (Karczewska et al., 2021; Mirhoseini et al., 2021; Schmutz et al., 2009). In contrast, factor 4 was not significantly correlated with workload according to the linear mixed model analysis, whereas Kendall's Tau showed a small but significant positive correlation. The lack of a relationship between factor 4 and workload is in line with a current review by Banu et al. (2021). According to the authors, there are many factors which influence the workload in a human-computer-interaction, but "Perceived privacy and security" or similar concepts are not among them. In addition, the small positive correlation could be explained by the concept of information overload. This concept describes a state in which a person is presented with too much information, causing high cognitive demands, stress and decreased performance (Eppler & Mengis, 2004). From this, it can be concluded that processing more information requires more effort from the person, leading to an increased workload. In the context of factor 4, high scores mean that the chatbot presented information on privacy issues to the participant. Thus, this additional information may have increased the cognitive demand associated with the interaction, explaining the positive correlation between factor 4 and workload. To summarize, this study found an expected negative relationship between workload and the BUS and its factors except for factor 4. This exception can be explained by the content of factor 4, which may not influence workload or have a positive relationship to workload due to the presentation of additional information.

The interpretation of this study's findings must also consider its limitations. Some of these limitations relate to the sampling of chatbots and respondents. To begin with, this study only included chatbots with high satisfaction and low workload rates (Table 2). Consequently, the findings of this study may not be generalizable to chatbots with a high workload and low satisfaction rates. Furthermore, the sample size may not have been large enough to confidently interpret Chi-square in the confirmatory factor analysis, as Li (2016) only obtained accurate Chi-square values with a larger sample size than the one in this study. In

addition, non-random sampling methods, including asking acquaintances, promotion on social media and survey sharing websites, and recruiting students who are required to participate in studies, were used to find a relatively large number of participants in a short period of time. Since simple random samples are required to generalize findings to a population (Hirschauer et al., 2021), this means that the current findings may not apply to the general population of chatbot users. However, the findings likely apply at least to students in higher education, who were most likely to participate due to the recruitment procedures. Overall, the limitations related to sampling mean that this study is most applicable to chatbots with high satisfaction and low workload rates and likely to university students.

Moreover, there were limitations related to other aspects of this study. For example, there was only one task per chatbot. Therefore, it is possible that some participants judged their experience with the task rather than the chatbot. Likely, this was not an issue, as the instructions were written to prevent this problem and there were no indications of this problem in the data. Still, this possibility cannot irrefutably be excluded. Additionally, this study was conducted online, meaning that the participants' experiences were not standardized. For example, the use of different devices, different internet connections or background noise may have influenced the participants' ratings. While in the linear mixed model analyses, this was controlled for by including a random effect for the participants, this was not possible in the other analyses. Lastly, there was a high rate of incomplete responses only about half of the respondents completed the survey for all five chatbots. This may raise concerns about the seriousness of the respondents. Still, this was likely not an issue for two reasons. First, the findings largely complied with expectations, which would be unlikely with random, unserious answers. Second, the outcomes were highly similar when running the analyses with only those participants who completed the entire survey. All in all, limitations that were not related to sampling likely did not have a large influence on the findings.

Based on the limitations and findings of this study, specific directions for future research can be discerned. One recommendation is to clarify whether item 1 presents a Heywood case and if so, what causes it. Based on this information, item 1 may be removed or modified and the factor structure of the adapted BUS can be investigated. To add to that, studies are needed that avoid the limitations of this study to see how robust the current findings are. Specifically, this requires studies with multiple tasks per chatbot, large random samples, standardized settings, and more chatbots with low satisfaction rates and high workload scores. Beyond that, future research may extend the validation of the BUS. For this, researchers may investigate the relationships between the BUS and variables other than

workload to contribute to the establishment of a nomological network, a key process in scale validation (Rauvola et al., 2020). In addition, research may aim to validate the BUS factors, so they might be used to measure individual aspects of chatbot satisfaction in the future. In summary, future research should investigate the possibility of a Heywood case in item 1, replicate this study while avoiding its limitations, and add to the establishment of a nomological network for the BUS.

Conclusion

On the whole, the findings on the properties of the BUS are in line with expectations. This study validated the BUS as a measure of chatbot satisfaction, as it correlates with the UMUX-Lite. Furthermore, the BUS is reliable and the five-factor structure observed by Borsci et al. (2021a) was confirmed. Moreover, the BUS is related to workload as expected, lending further credibility to its measurements. Based on these findings, the BUS is a promising scale which can be used to measure overall chatbot satisfaction. On the other hand, item 1 has a low or possibly negative variance, which warrants further investigation. Future studies may focus on improving the generalizability of these findings, identifying whether item 1 is a Heywood case and if it needs to be modified or removed, and on studying the relationships between the BUS and other variables. In conclusion, the BUS can be confirmed as reliable, having a five-factor structure, correlating with workload and measuring chatbot satisfaction.

These findings are relevant to improving the quality of chatbots by facilitating the use of the BUS as a standardized measure of chatbot satisfaction. As chatbots are a frequently used technology (Behera et al., 2021; Caldarini et al., 2022; Hwang et al., 2020; Lasek & Jessa, 2013; Quan et al., 2019) and offer many benefits to both businesses and end users (Behera et al., 2021), they should be designed in a way that satisfies users so they may be adopted successfully. For this, a standardized measure for chatbot satisfaction is needed, since this will allow researchers and developers to compare chatbots, replicate studies, and find benchmarks for desired satisfaction levels. Further, this scale needs to be specific to chatbots, as the findings of this study, Borsci et al. (2021b) and Waldmann (2021) show that chatbot satisfaction encompasses more aspects than satisfaction with many other systems. In this study, the BUS was validated as a reliable measure of this specific construct. Therefore, the present findings support the use of the BUS as a standardized chatbot satisfaction scale and thus, contribute to the improvement of chatbots and the development of best practices for their design.

References

- Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, Article 100006. <https://doi.org/10.1016/j.mlwa.2020.100006>
- Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, 54, Article 101473. <https://doi.org/10.1016/j.tele.2020.101473>
- Banu, R., Al Siyabi, W. S. A., & Al Minje, Y. (2021). A conceptual review on integration of cognitive load theory and human-computer interaction. In *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)* (pp. 667-672). <https://doi.org/10.1109/ICSECS52883.2021.00127>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Behera, R. K., Bala, P. K., & Ray, A. (2021). Cognitive chatbot for personalised contextual customer service: Behind the scene and beyond the hype. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-021-10168-y>
- Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, 31(8), 484-495. <https://doi.org/10.1080/10447318.2015.1064648>
- Borsci, S., Buckle, P., & Walne, S. (2020). Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? [Article]. *Applied Ergonomics*, 84, Article 103007. <https://doi.org/10.1016/j.apergo.2019.103007>
- Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A. & van der Velde, F. (2021a) *Confirmatory factorial analysis of the chatbot usability scale: A multilanguage validation* [Manuscript submitted for publication]. Faculty of Behavioural, Management and Social Sciences (BMS), University of Twente
- Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2021b). 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>
- Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: Changing user needs and motivations. *Interactions*, 25(5), 38-43. <https://doi.org/10.1145/3236669>
- Bryant, F. B., Yarnold, P. R., & Michelson, E. A. (1999). Statistical methodology: VIII. Using confirmatory factor analysis (CFA) in emergency medicine research. *Academic Emergency Medicine*, 6(1), 54-66. <https://doi.org/10.1111/j.1553-2712.1999.tb00096.x>
- Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information (Switzerland)*, 13(1), Article 41. <https://doi.org/10.3390/info13010041>
- Cheng, Y., & Jiang, H. (2021). Customer–brand relationship in the era of artificial intelligence: Understanding the role of chatbot marketing efforts. *Journal of Product & Brand Management*, 31(2), 252-264. <https://doi.org/10.1108/jpbm-05-2020-2907>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <https://doi.org/10.1037/h0040957>
- Devos, H., Gustafson, K., Ahmadnezhad, P., Liao, K., Mahnken, J. D., Brooks, W. M., & Burns, J. M. (2020). Psychometric properties of NASA-TLX and Index of Cognitive Activity as measures of cognitive workload in older adults. *Brain Sciences*, 10(12), 994. <https://doi.org/10.3390/brainsci10120994>
- Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society*, 20(5), 325-344. <https://doi.org/10.1080/01972240490507974>
- Epskamp, S., Stuber, S., Nak, J., Veenamn, M., & Jorgnesen, T. D. (2022). *SemPlot: Path diagrams and visual analysis of various SEM packages' output*. The Comprehensive R Archive Network (CRAN). <https://CRAN.R-project.org/package=semPlot>
- Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97, 304-316. <https://doi.org/10.1016/j.chb.2019.01.020>
- Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-11). <https://doi.org/10.1145/3290605.3300439>

- Gupta, A., Hahthwar, D., & Vijayakumar, A. (2020). Introduction to AI chatbots. *International Journal of Engineering Research & Technology*, 9(7), 255-258. <https://doi.org/10.17577/IJERTV9IS070143>
- Hart, S. G., & Wickens, C. D. (1990). Workload assessment and prediction. In H. R. Booher (Ed.), *Manprint: An approach to system integration* (pp. 25-296). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-0437-8_9
- Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C., & Jantsch, A. (2021). Inference using non-random samples? Stop right there! *Significance*, 18(5), 20 - 24. <https://doi.org/10.1111/1740-9713.01568>
- Hwang, T. H., Lee, J., Hyun, S. M., & Lee, K. (2020). Implementation of interactive healthcare advisor model using chatbot and visualization. In *International Conference on ICT Convergence* (pp. 452-455). <https://doi.org/10.1109/ICTC4980.2020.9289621>
- International Organization for Standardization. (2018). *Ergonomics of Human-System Interaction - Part 11: Usability: Definitions and Concepts* (ISO Standard No. 9241-11).
- Kantowitz, B. H. (1987). 3. Mental workload. In P. A. Hancock (Ed.), *Human factors psychology* (pp. 81-121). Elsevier B. V. [https://doi.org/10.1016/s0166-4115\(08\)62307-9](https://doi.org/10.1016/s0166-4115(08)62307-9)
- Karczewska, B., Kukla, E., Muke, P. Z., Telec, Z., & Trawiński, B. (2021). Usability study of mobile applications with cognitive load resulting from environmental factors. In N. T. Nguyen, S. Chittayasothorn, D. Niyato, & B. Trawinsik (Eds.), *ACIIDS 2021: Intelligent Information and Database Systems* (pp. 851-864). Springer, Cham. https://doi.org/10.1007/978-3-030-73280-6_67
- Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)* (pp. 555-565). Association for Computing Machinery, New York. <https://doi.org/10.1145/3064663.3064672>
- Kull, A. J., Romero, M., & Monahan, L. (2021). How may I help you? Driving brand engagement through the warmth of an initial chatbot message. *Journal of Business Research*, 135, 840-850. <https://doi.org/10.1016/j.jbusres.2021.03.005>
- Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>.

- Lamm, K., Lamm, A., & Edgar, D. (2020). Scale development and validation: Methodology and recommendations. *Journal of International Agricultural and Extension Education*, 27(2), 24-35. <https://doi.org/10.5191/jiaee.2020.27224>
- Lasek, M., & Jessa, S. (2013). Chatbots for customer service on hotels' websites. *Information Systems in Management*, 2(2), 146 - 158.
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 2099-2102). Association for Computing Machinery, New York. <https://doi.org/10.1145/2470654.2481287>
- Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936-949. <https://doi.org/10.3758/s13428-015-0619-7>
- Liu, M., Ding, Q., Zhang, Y., Zhao, G., Hu, C., Gong, J., Xu, P., Zhang, Y., Zhang, L., & Wang, Q. (2020). Cold comfort matters - How channel-wise emotional strategies help in a customer service chatbot. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)* (pp. 1-7). Association for Computing Machinery, New York. <https://doi.org/10.1145/3334480.3382905>
- Lopez, S. M. K. (2021). *Confirmatory factor analysis of a new satisfaction scale for conversational agents and the role of decision-making styles* [Unpublished bachelor's thesis, University of Twente]. University of Twente student theses. <https://essay.utwente.nl/86852/>
- Mirhoseini, M., Pagé, S.-A., Léger, P.-M., & Sénécal, S. (2021). What deters online grocery shopping? Investigating the effect of arithmetic complexity and product type on user satisfaction. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(4), 828-845. <https://doi.org/10.3390/jtaer16040047>
- National Aeronautics and Space Administration (n.d.). *NASA task load index (TLX). Pencil and paper package*. https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX_pappen_manual.pdf
- Nguyen, D. M., Chiu, Y. T. H., & Le, H. D. (2021). Determinants of continuance intention towards banks' chatbot services in vietnam: A necessity for sustainable development. *Sustainability*, 13(14). <https://doi.org/10.3390/su13147625>
- Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. menu-based interfaces: An empirical study. *Computers in Human Behavior*, 128, Article 107093. <https://doi.org/10.1016/j.chb.2021.107093>

- Nicolescu, L., & Tudorache, M. T. (2022). Human-computer interaction in customer service: The experience with AI chatbots – A systematic literature review. *Electronics*, *11*(10), Article 1579. <https://doi.org/10.3390/electronics11101579>
- Quan, T., Trinh, T., Ngo, D., Pham, H., Hoang, L., Hoang, H., Thai, T., Vo, P., Pham, D., & Mai, T. (2019). Lead engagement by automated real estate chatbot. In *NICS 2018 - Proceedings of 2018 5th NAFOSTED Conference on Information and Computer Science* (pp. 357-359). <https://doi.org/10.1109/NICS.2018.8606862>
- Qualtrics (2005). *Qualtrics XM* [Computer Software]. <https://www.qualtrics.com>
- R Core Team (2022). *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- Rauvola, R. S., Briggs, E. P., & Hinyard, L. J. (2020). Nomology, validity, and interprofessional research: The missing link(s). *Journal of Interprofessional Care*, *34*(4), 545-556. <https://doi.org/10.1080/13561820.2020.1712333>
- Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1-25. <https://doi.org/10.18637/jss.v017.i05>
- Rosseel, Y. (2012). Lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Ruiz-Rabelo, J. F., Navarro-Rodriguez, E., Di-Stasi, L. L., Diaz-Jimenez, N., Cabrera-Bermon, J., Diaz-Iglesias, C., Gomez-Alvarez, M., & Briceno-Delgado, J. (2015). Validation of the NASA-TLX score in ongoing assessment of mental workload during a laparoscopic learning curve in bariatric surgery. *Obesity Surgery*, *25*(12), 2451-2456. <https://doi.org/10.1007/s11695-015-1922-1>
- Said, S., Gozdzik, M., Roche, T. R., Braun, J., Rossler, J., Kaserer, A., Spahn, D. R., Nothiger, C. B., & Tscholl, D. W. (2020). Validation of the raw national aeronautics and space administration task load index (NASA-TLX) questionnaire to assess perceived workload in patient monitoring tasks: Pooled analysis study using mixed models. *Journal of Medical Internet Research*, *22*(9), e19472. <https://doi.org/10.2196/19472>
- Schmutz, P., Heinz, S., Métrailler, Y., & Opwis, K. (2009). Cognitive load in eCommerce applications - Measurement and effects on user satisfaction. *Advances in Human-Computer Interaction*, *2009*, 1-9. <https://doi.org/10.1155/2009/121494>

- Tran, A. D., Pallant, J. I., & Johnson, L. W. (2021). Exploring the impact of chatbots on consumer sentiment and expectations in retail. *Journal of Retailing and Consumer Services*, 63, Article 102718. <https://doi.org/10.1016/j.jretconser.2021.102718>
- Van den Bos, M. (2021). *Testing a scale for perceived usability and user satisfaction in chatbots: Testing the BotScale* [Unpublished master's thesis, University of Twente]. University of Twente student theses. <https://essay.utwente.nl/85767/>
- Waldmann, A. (2021). *User satisfaction and trust in chatbots: Testing the chatbot usability scale and the relationship of trust and satisfaction in the interaction with chatbots* [Unpublished bachelor's thesis, University of Twente]. University of Twente student theses. <https://essay.utwente.nl/87443/>
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45. <https://doi.org/10.1145/365153.365168>
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis (Use R)*. Springer-Verlag New York.
- Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. *Zhonghua lao dong wei sheng zhi ye bing za zhi = Zhonghua laodong weisheng zhiyebing zazhi = Chinese journal of industrial hygiene and occupational diseases*, 23(3), 178-181. <https://doi.org/10.360/CMA.J.ISSN.1001.9391.2005.03.007>
- Zhou, Y. (2019). A mixed methods model of scale development and validation analysis. *Measurement*, 17(1), 38-47. <https://doi.org/10.1080/15366367.2018.1479088>

Appendix A

Demographic Questions

To begin, please provide some demographic information about yourself. This data will not be used to identify you personally. It is only used to describe the sample of people that take part in this study.

What is your age in years?

What is your level of proficiency in English?

- Basic (A1 - A2)
- Intermediate (B1 - B2+)
- Advanced (C1 - C2)
- Native speaker

What is your sex as assigned at birth?

- Male
- Female

Please indicate to what extent you agree with the following statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I am familiar with chatbots and/or other conversational interfaces.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I know how chatbots work.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel confident with using chatbots.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix B

Chatbots and Task Scenarios

Table A1

Chatbots with Company or Institution, Weblink and Task Scenario

Company or Institution	Weblink	Task Scenario
Samsung	https://www.samsung.com/us/	“You live in the USA and have ordered a TV by Samsung. However, the image is flickering, so you want to chat with Samsung’s support. Once you have found their chatbot, use it to schedule a repair for your flickering TV. Your task is done once the chatbot gives you the option to schedule a repair, so you can stop without really making an appointment.”
State of Mississippi	https://www.ms.gov/home	“You live in Mississippi and have recently lost your job. You want to use their chatbot to find out how to file an unemployment claim and whether the government can help you to find a new job.”
Virtual Spirits	https://www.virtualspirits.com/	“You work for a large company and are asked to automate their customer service with a chatbot. For this, you are considering Virtual Spirits, a company that sells chatbots to other businesses. Find out how the customer service chatbot sold by Virtual Spirits works and what pricing plans they offer. Then, you decide you want to ask further questions


to a human, so find out how to get in touch with a human employee from Virtual Spirits by email.

You do not need to really get in touch with a human, so your task is done once the chatbot gives you the option to contact or be contacted by an employee.”

Peloton <https://www.onepeloton.com/> “You live in the USA and have an all-access membership for Peloton, but you decided you do not need it anymore, so you want to cancel it. Use Peloton’s chatbot to get support and find out how to cancel your membership.”

Message Envy <https://www.messageenvy.com/memberships?membership-drive-modal> “You live in Chicago (zip code 60610) and like to regularly get a massage at Message Envy. So, you are considering getting a membership to save money. For this, you want to find out how much a 60-minute or 90-minute massage session costs with or without a membership at the location nearest to you.”

Zoom <https://zoom.us/> “Imagine you are in a Zoom meeting with a friend, preparing your homework together. You start experimenting with the settings in Zoom and remember seeing different backgrounds on other people when they use Zoom. Unfortunately, your friend also does not know how to change the background. So, you decide to get help on this via the

Zoom website. Now, your task is to find the chatbot and ask for help. Please open the official Zoom website and look for the chatbot function. Often it is a chat symbol popping up in the corner like the one you can see here . Then, try finding needed information through the suggestions that the chatbot provides you with. After finding the video with the instruction, your task is finished.”

Kia

<https://www.kia.com/uk/>

“You live in London (postal code SE10 8BL) and you want to buy a new car from Kia. So, you use Kia’s chatbot to find a Kia store in your area. You also want to get this store’s phone number and opening times.”

Appendix C

BUS-Scale with Instructions

Respond to the next statements based on your experience with the chatbot.

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

Appendix E

Adapted NASA-TLX with Instructions

Please read the definitions of the factors that influence workload. Then move the sliders to rate the factors in your experience with the chatbot, not the task itself.

Title	Endpoints	Descriptions
Mental Demand	Low/High	How much mental and perceptual activity, was required (eg., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low/High	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low/High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Effort	Low/High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low/High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
Performance	Good/Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

How much of these factors did it take to solve the task with the help of the chatbot? Please rate your experience with the chatbot, not the task itself.

Low High
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Mental Demand



Physical Demand



Temporal Demand



Effort



Frustration



How would you rate your own performance in solving the task with the help of the chatbot? Please rate your experience with the chatbot, not the task itself.

Good Poor
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Performance



Appendix F

Study Information and Informed Consent

This study aims to contribute to the validation of the new BUS-scale, standardized scale for chatbot satisfaction, and to investigate the relationship between mental workload and chatbot satisfaction.

For this, you are asked to interact with five customer service chatbots by solving specific tasks and to fill in the BUS-Scale and the NASA-TLX, a questionnaire measuring mental workload, for each chatbot. In all, this usually takes about 30 minutes.

Your data will be used and analysed to write bachelor theses on the ongoing validation and development of the BUS-Scale. The data will be anonymised, meaning that it will not be tied to your name or other personal information that could identify you. The anonymised data may be seen by student researchers and the supervisor of the project.

We expect that this study poses no burdens or risks to you. Still, you can withdraw from the study at any time without needing to justify your withdrawal. If you have any complaints or questions, please contact the researcher listed below. This research has been approved by the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente.

In case of complaints or questions, please contact:

Mustafa Taha

m.a.taha@student.utwente.nl

I have read and understood the study information. I voluntarily consent to participate in this study under the terms stated in the study information.

Yes

No

Appendix G

General Instructions

In the following, you will be presented with five chatbots and a scenario for each. For each chatbot, read the scenario, then click the link to find the chatbot and use it to fulfil the task specified in the scenario.

The task is done once the chatbot offers you the necessary information or gives you a direct link to the information. You do not need to click the links and read the information. Similarly, you do not need to really schedule appointments, repairs, or to really contact human employees. It is enough if the chatbot gives you the option to do so. Do not give the chatbots any personal information. This is not necessary to solve the tasks.

After finishing the task, proceed with the survey and fill in the questionnaires for the first chatbot. Then, you will be directed to the next chatbot and scenario.

If you do not finish the task within 10 minutes, please proceed with the survey.

Appendix H

R Script for Data Analysis

```
#Install packages#  
  
#install.packages("tidyverse")  
  
#install.packages("psych")  
  
#install.packages("dplyr")  
  
#install.packages("purrr")  
  
#install.packages("ggplot2")  
  
#install.packages("haven")  
  
#install.packages("CTT")  
  
#install.packages("Lambda4")  
  
#install.packages("mirt")  
  
#install.packages("janitor")  
  
#install.packages("broom")  
  
#install.packages("lavaan")  
  
#install.packages("lme4")  
  
#install.packages("nlme")  
  
#install.packages("lmerTest")  
  
#install.packages("summarytools")  
  
#install.packages("readxl")  
  
#install.packages("foreign")  
  
#install.packages("lavaanPlot")  
  
#install.packages("ggpubr")  
  
#install.packages("knitr")
```

```
#install.packages("semPlot")  
  
#install.packages("MVN")  
  
#install.packages("tidyr")  
  
#install.packages("WriteXLS")  
  
#install.packages("ltm")  
  
#install.packages("outliers")  
  
#install.packages("EnvStats")  
  
  
#Load packages  
  
library(tidyverse)  
  
library(psych)  
  
library(dplyr)  
  
library(purrr)  
  
library(ggplot2)  
  
library(haven)  
  
library(CTT)  
  
library(Lambda4)  
  
library(mirt)  
  
library(janitor)  
  
library(broom)  
  
library(lavaan) #confirmatory factor analysis  
  
library(lme4) #linear mixed model  
  
library(nlme) #linear mixed model  
  
library(lmerTest) #p-values for linear mixed model
```

```
library(summarytools)

library(readxl)

library(foreign)

library(lavaanPlot)

library(ggpubr)

library(knitr)

library(semPlot)

library(MVN)

library(tidyr)

library(WriteXLS)

library(ltm)

library(outliers)

library(EnvStats)

#Import file

setwd("C:/Users/matah/Desktop/University/Module 12 Mustafa/Data")

raw_data <- read.csv("Data V13 numeric values - Mustafa's AND Marias data.csv")

#remove unnecessary columns and rows

data_with_question_text <- raw_data %>% select(12:158)

data_with_question_text <- subset(raw_data, select = -c(1:11))

data <- data_with_question_text[-c(1,2), ]

#renaming demographics
```

```
data <- rename(data, age = Q53)

data <- rename(data, English_proficiency = Q115)

data <- rename(data, sex = Q54)

#frequency tables to check for missing data

data$age %>% map(tabyl)

#creating chatbot experience variable for demographics

#used near the end of the R Script

chatbot_experience <- (as.numeric(data$Q82_1) + as.numeric(data$Q82_1) +
as.numeric(data$Q82_1)) / 3

####PREPARING DATA SET FOR ANALYSES#####

#creating Participant variable

no_participants <- seq_along(data$age)

participants <- vector()

for (i in no_participants) {participants <- c(participants, i)}

participants <- c(participants, participants, participants, participants, participants,
participants)

#creating chatbot variable

chatbot <- vector()

for (i in no_participants) {chatbot <- c(chatbot, "Samsung")}

for (i in no_participants) {chatbot <- c(chatbot, "Government of Mississippi)}
```

```
for (i in no_participants) {chatbot <- c(chatbot, "Virtual Spirits")}
```

```
for (i in no_participants) {chatbot <- c(chatbot, "Peloton")}
```

```
for (i in no_participants) {chatbot <- c(chatbot, "Message Envy")}
```

```
for (i in no_participants) {chatbot <- c(chatbot, "Zoom")}
```

```
for (i in no_participants) {chatbot <- c(chatbot, "Kia")}
```

```
#creating BUS item variables
```

```
Item_1 <- as.numeric(c(data$Q62_1, data$Q63_1, data$Q64_1, data$Q65_1, data$Q66_1,  
data$Q67_1, data$Q68_1))
```

```
Item_2 <- as.numeric(c(data$Q62_2, data$Q63_2, data$Q64_2, data$Q65_2, data$Q66_2,  
data$Q67_2, data$Q68_2))
```

```
Item_3 <- as.numeric(c(data$Q62_3, data$Q63_3, data$Q64_3, data$Q65_3, data$Q66_3,  
data$Q67_3, data$Q68_3))
```

```
Item_4 <- as.numeric(c(data$Q62_4, data$Q63_4, data$Q64_4, data$Q65_4, data$Q66_4,  
data$Q67_4, data$Q68_4))
```

```
Item_5 <- as.numeric(c(data$Q62_5, data$Q63_5, data$Q64_5, data$Q65_5, data$Q66_5,  
data$Q67_5, data$Q68_5))
```

```
Item_6 <- as.numeric(c(data$Q62_6, data$Q63_6, data$Q64_6, data$Q65_6, data$Q66_6,  
data$Q67_6, data$Q68_6))
```

```
Item_7 <- as.numeric(c(data$Q62_7, data$Q63_7, data$Q64_7, data$Q65_7, data$Q66_7,  
data$Q67_7, data$Q68_7))
```

```
Item_8 <- as.numeric(c(data$Q62_8, data$Q63_8, data$Q64_8, data$Q65_8, data$Q66_8,  
data$Q67_8, data$Q68_8))
```

```
Item_9 <- as.numeric(c(data$Q62_9, data$Q63_9, data$Q64_9, data$Q65_9, data$Q66_9,  
data$Q67_9, data$Q68_9))
```

```
Item_10 <- as.numeric(c(data$Q62_10, data$Q63_10, data$Q64_10, data$Q65_10,  
data$Q66_10, data$Q67_10, data$Q68_10))
```

```
Item_11 <- as.numeric(c(data$Q62_11, data$Q63_11, data$Q64_11, data$Q65_11,  
data$Q66_11, data$Q67_11, data$Q68_11))
```

```
#creating UMUX_Item variables
```

```
UMUX_requirements <- as.numeric(c(data$Q84_1, data$Q85_1, data$Q86_1, data$Q87_1,  
data$Q88_1, data$Q90_1, data$Q89_1))
```

```
UMUX_easy <- as.numeric(c(data$Q84_2, data$Q85_2, data$Q86_2, data$Q87_2,  
data$Q88_2, data$Q90_2, data$Q89_2))
```

```
#creating NASA-TLX dimension variables
```

```
NASA_mental_demand <- as.numeric(c(data$Q9_1, data$Q93_1, data$Q97_1, data$Q101_1,  
data$Q105_1, data$Q113_1, data$Q109_1))
```

```
NASA_physical_demand <- as.numeric(c(data$Q9_2, data$Q93_2, data$Q97_2,  
data$Q101_2, data$Q105_2, data$Q113_2, data$Q109_2))
```

```
NASA_temporal_demand <- as.numeric(c(data$Q9_3, data$Q93_3, data$Q97_3,  
data$Q101_3, data$Q105_3, data$Q113_3, data$Q109_3))
```

```
NASA_effort <- as.numeric(c(data$Q9_4, data$Q93_4, data$Q97_4, data$Q101_4,  
data$Q105_4, data$Q113_4, data$Q109_4))
```

```
NASA_frustration <- as.numeric(c(data$Q9_5, data$Q93_5, data$Q97_5, data$Q101_5,  
data$Q105_5, data$Q113_5, data$Q109_5))
```

```
NASA_performance <- as.numeric(c(data$Q55_1, data$Q94_1, data$Q98_1, data$Q102_1,  
data$Q106_1, data$Q114_1, data$Q110_1))
```

```
#compiling item variables in a data frame
```

```
new_data <- data.frame(participants, chatbot, Item_1, Item_2, Item_3, Item_4, Item_5,  
Item_6, Item_7, Item_8, Item_9, Item_10, Item_11, UMUX_requirements, UMUX_easy,  
NASA_mental_demand, NASA_physical_demand, NASA_temporal_demand, NASA_effort,  
NASA_frustration, NASA_performance)
```

```
new_data <- subset(new_data, UMUX_easy == 1 | UMUX_easy == 2 | UMUX_easy == 3 |
  UMUX_easy == 4 | UMUX_easy == 5 | UMUX_easy == 6 | UMUX_easy == 7)
```

```
new_data[c('participants', 'chatbot')] %>% map(tabyl) # check number of observations per
  participant and chatbot
```

```
#calculating overall scores and adding them to the data frame
```

```
BUS_score <- (((new_data$Item_1) + (new_data$Item_2) + (new_data$Item_3) +
  (new_data$Item_4) + (new_data$Item_5) + (new_data$Item_6) + (new_data$Item_7) +
  (new_data$Item_8) + (new_data$Item_9) + (new_data$Item_10) + (new_data$Item_11)) /
  11) / 5
```

```
BUS_score <- na.omit(BUS_score)
```

```
new_data$BUS_score <- BUS_score
```

```
BUS_1 <- (((new_data$Item_1) + (new_data$Item_2)) / 2) / 5
```

```
BUS_1 <- na.omit(BUS_1)
```

```
new_data$BUS_1 <- BUS_1
```

```
BUS_2 <- (((new_data$Item_3) + (new_data$Item_4) + (new_data$Item_5)) / 3) / 5
```

```
BUS_2 <- na.omit(BUS_2)
```

```
new_data$BUS_2 <- BUS_2
```

```
BUS_3 <- (((new_data$Item_6) + (new_data$Item_7) + (new_data$Item_8) +
  (new_data$Item_9)) / 4) / 5
```

```
BUS_3 <- na.omit(BUS_3)
```

```
new_data$BUS_3 <- BUS_3
```

```
BUS_4 <- (new_data$Item_10) / 5
```

```
BUS_4 <- na.omit(BUS_4)
```

```
new_data$BUS_4 <- BUS_4
```

```
BUS_5 <- (new_data$Item_11) / 5
```

```
BUS_5 <- na.omit(BUS_5)
```

```
new_data$BUS_5 <- BUS_5
```

```
UMUX_Lite <- (((new_data$UMUX_easy) + (new_data$UMUX_requirements)) / 2) / 7
```

```
new_data$UMUX_Lite <- UMUX_Lite
```

```
NASA_workload <- (new_data$NASA_mental_demand +
new_data$NASA_physical_demand + new_data$NASA_temporal_demand +
new_data$NASA_effort + new_data$NASA_frustration + new_data$NASA_performance) /
6 / 100
```

```
new_data$NASA_workload <- NASA_workload
```

```
summary(new_data) ##descriptives of all new variables to check if everything is fine
```

```
#####
```

```
###CONFIRMATORY FACTOR ANALYSIS###
```

```
#Normality check
```

```
shapiro.test(new_data$Item_1)
```

```
shapiro.test(new_data$Item_2)
```



```
shapiro.test(new_data$Item_3)
shapiro.test(new_data$Item_4)
shapiro.test(new_data$Item_5)
shapiro.test(new_data$Item_6)
shapiro.test(new_data$Item_7)
shapiro.test(new_data$Item_8)
shapiro.test(new_data$Item_9)
shapiro.test(new_data$Item_10)
shapiro.test(new_data$Item_11)
```

```
#specifying factor model
```

```
CFA_model <- '
```

```
F1 =~ Item_1 + Item_2
```

```
F2 =~ Item_3 + Item_4 + Item_5
```

```
F3 =~ Item_6 + Item_7 + Item_8 + Item_9
```

```
F4 =~ Item_10
```

```
F5 =~ Item_11'
```

```
#F1 Perceived accessibility to chatbot functions
```

```
#F2 Perceived quality of chatbot functions
```

```
#F3 Perceived quality of conversation and information provided
```

```
#F4 Perceived privacy and security
```

```
#F5 Time response
```

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

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

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

```
#####
```

```
###RELIABILITY
```

```
#BUS 11
```

```
alphaBUS11E3 <-data.frame(new_data$Item_1, new_data$Item_2, new_data$Item_3,
new_data$Item_4, new_data$Item_5, new_data$Item_6, new_data$Item_7,
new_data$Item_8, new_data$Item_9, new_data$Item_10, new_data$Item_11)
```

```
cronbach.alpha(alphaBUS11E3)
```

```
#F1
```

```
alphaF1E3<-data.frame(new_data$Item_1, new_data$Item_2)
```

```
cronbach.alpha(alphaF1E3, standardized = TRUE, CI = TRUE)
```

```
#F2
```

```
alphaF2E3 <-data.frame(new_data$Item_3, new_data$Item_4, new_data$Item_5)
```

```
cronbach.alpha(alphaF2E3, standardized = TRUE, CI = TRUE)
```

```
#F3
```

```
alphaF3E3 <-data.frame(new_data$Item_6, new_data$Item_7, new_data$Item_8,
new_data$Item_9)
```

```
cronbach.alpha(alphaF3E3, standardized = TRUE, CI = TRUE)
```

```
#####
```

```
#Filtering out data without NASA-TLX responses and with UMUX-Lite on a 5-point scale
instead of 7-points
```

```
new_data <- na.omit(new_data)
```

```
#####
```

```
###RELATION BETWEEN BUS AND UMUX-LITE###
```

```
##Kendall's Tau BUS11
```

```
shapiro.test(new_data$BUS_score)
```

```
shapiro.test(new_data$UMUX_Lite)
```

```
cor.test(new_data$BUS_score, new_data$UMUX_Lite,use="pairwise.complete.obs", method
= "kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=UMUX_Lite, y=BUS_score)) +
```

```
  xlab("UMUX Lite overall scores") + ylab("BUS11 overall scores")+
```

```
  geom_point() +
```

```
  geom_smooth(method = "lm",se = F, fullrange = F)
```

```
##Kendall's Tau F1
```

```
shapiro.test(new_data$BUS_1)
```

```
shapiro.test(new_data$UMUX_Lite)
```

```
cor.test(new_data$BUS_1, new_data$UMUX_Lite,use="pairwise.complete.obs", method =  
"kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=UMUX_Lite, y=BUS_1)) +  
  xlab("UMUX Lite overall scores") + ylab("BUS factor 1")+  
  geom_point() +  
  geom_smooth(method = "lm",se = F, fullrange = F)
```

```
##Kendall's Tau F2
```

```
shapiro.test(new_data$BUS_2)
```

```
shapiro.test(new_data$UMUX_Lite)
```

```
cor.test(new_data$BUS_2, new_data$UMUX_Lite,use="pairwise.complete.obs", method =  
"kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=UMUX_Lite, y=BUS_2)) +  
  xlab("UMUX Lite overall scores") + ylab("BUS factor 2")+  
  geom_point() +  
  geom_smooth(method = "lm",se = F, fullrange = F)
```

```
##Kendall's Tau F3
```

```
shapiro.test(new_data$BUS_3)
```

```
shapiro.test(new_data$UMUX_Lite)
```

```
cor.test(new_data$BUS_3, new_data$UMUX_Lite,use="pairwise.complete.obs", method =  
"kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=UMUX_Lite, y=BUS_3)) +  
  xlab("UMUX Lite overall scores") + ylab("BUS factor 3")+  
  geom_point() +  
  geom_smooth(method = "lm",se = F, fullrange = F)
```

```
##Kendall's Tau F4
```

```
shapiro.test(new_data$BUS_4)
```

```
shapiro.test(new_data$UMUX_Lite)
```

```
cor.test(new_data$BUS_4, new_data$UMUX_Lite,use="pairwise.complete.obs", method =  
"kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=UMUX_Lite, y=BUS_4)) +  
  xlab("UMUX Lite overall scores") + ylab("BUS factor 4")+  
  geom_point() +
```

```
geom_smooth(method = "lm",se = F, fullrange = F)

##Kendall's Tau F5

shapiro.test(new_data$BUS_5)

shapiro.test(new_data$UMUX_Lite)

cor.test(new_data$BUS_5, new_data$UMUX_Lite,use="pairwise.complete.obs", method =
"kendall")

#Plot of correlation

ggplot(new_data,aes(x=UMUX_Lite, y=BUS_5)) +
  xlab("UMUX Lite overall scores") + ylab("BUS factor 5")+
  geom_point() +
  geom_smooth(method = "lm",se = F, fullrange = F)

##Linear Mixed Model

#checking linearity of the data

plot(x = new_data$BUS_score, y = new_data$UMUX_Lite)

plot(x = new_data$BUS_1, y = new_data$UMUX_Lite)

plot(x = new_data$BUS_2, y = new_data$UMUX_Lite)

plot(x = new_data$BUS_3, y = new_data$UMUX_Lite)

plot(x = new_data$BUS_4, y = new_data$UMUX_Lite)

plot(x = new_data$BUS_5, y = new_data$UMUX_Lite)

#checking for clustering between UMUX_Lite and participants
```

```
boxplot(UMUX_Lite ~ participants, data = new_data, xlab = "participants", #boxplot
        ylab = "UMUX-Lite", main = "Clustering in UMUX-Lite scores by participants")

#checking for clustering between UMUX_Lite and chatbots
boxplot(UMUX_Lite ~ chatbot, data = new_data, xlab = "chatbot",      #boxplot
        ylab = "UMUX_Lite", main = "Clustering in UMUX-Lite scores by chatbots")

##models for relationship with UMUX_Lite - choose the one that is applicable
##1: random effect participants
#lmm_UMUX <- lmer(UMUX_Lite ~ BUS_score + (1|participants), data = new_data)

#check distribution of residuals
#res_UMUX1 <- residuals(lmm_UMUX)
#plot(fitted(lmm_UMUX), res_UMUX1) #homoscedasticity (equal variance) check
#abline(0,0)

#qqnorm(res_UMUX1) #normal distribution check - dots should match line
#qqline(res_UMUX1)

#plot(density(res_UMUX1)) #normal distribution check - should have bell curve centered on
0

#get output
#anova(lmm_UMUX)
#summary(lmm_UMUX)
```

```
##2: random effect chatbot
```

```
#lmm_UMUX2 <- lmer(UMUX_Lite ~ BUS_5 + (1|chatbot), data = new_data)
```

```
#check distribution of residuals
```

```
#res_UMUX2 <- residuals(lmm_UMUX2)
```

```
#plot(fitted(lmm_UMUX2), res_UMUX2) #homoscedascity (equal variance) check
```

```
#abline(0,0)
```

```
#qqnorm(res_UMUX2) #normal distribution check - dots should match line
```

```
#qqline(res_UMUX2)
```

```
#plot(density(res_UMUX2)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_UMUX2)
```

```
#summary(lmm_UMUX2)
```

```
##3: random effect participants and chatbot
```

```
###BUS-11
```

```
lmm_UMUX3 <- lmer(UMUX_Lite ~ BUS_score + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_UMUX3 <- residuals(lmm_UMUX3)
```



```
plot(fitted(lmm_UMUX3), res_UMUX3) #homoscedascity (equal variance) check
abline(0,0)

qqnorm(res_UMUX3) #normal distribution check - dots should match line
qqline(res_UMUX3)

plot(density(res_UMUX3)) #normal distribution check - should have bell curve centred on 0

#get output
#anova(lmm_UMUX3)
#info from anova are included in the summary() command
summary(lmm_UMUX3)

####F1
lmm_UMUX3.1 <- lmer(UMUX_Lite ~ BUS_1 + (1|participants) + (1|chatbot), data =
new_data)

#check distribution of residuals
res_UMUX3.1 <- residuals(lmm_UMUX3.1)
plot(fitted(lmm_UMUX3.1), res_UMUX3.1) #homoscedascity (equal variance) check
abline(0,0)

qqnorm(res_UMUX3.1) #normal distribution check - dots should match line
qqline(res_UMUX3.1)
```

```
plot(density(res_UMUX3.1)) #normal distribution check - should have bell curve centred on  
0
```

```
#get output
```

```
#anova(lmm_UMUX3.1)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_UMUX3.1)
```

```
###F2
```

```
lmm_UMUX3.2 <- lmer(UMUX_Lite ~ BUS_2 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_UMUX3.2 <- residuals(lmm_UMUX3.2)
```

```
plot(fitted(lmm_UMUX3.2), res_UMUX3.2) #homoscedascity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_UMUX3.2) #normal distribution check - dots should match line
```

```
qqline(res_UMUX3.2)
```

```
plot(density(res_UMUX3.2)) #normal distribution check - should have bell curve centred on  
0
```

```
#get output
```

```
#anova(lmm_UMUX3.2)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_UMUX3.2)
```

```
###F3
```

```
lmm_UMUX3.3 <- lmer(UMUX_Lite ~ BUS_3 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_UMUX3.3 <- residuals(lmm_UMUX3.3)
```

```
plot(fitted(lmm_UMUX3.3), res_UMUX3.3) #homoscedascity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_UMUX3.3) #normal distribution check - dots should match line
```

```
qqline(res_UMUX3.3)
```

```
plot(density(res_UMUX3.3)) #normal distribution check - should have bell curve centred on  
0
```

```
#get output
```

```
#anova(lmm_UMUX3.3)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_UMUX3.3)
```

```
###F4
```

```
lmm_UMUX3.4 <- lmer(UMUX_Lite ~ BUS_4 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_UMUX3.4 <- residuals(lmm_UMUX3.4)
```

```
plot(fitted(lmm_UMUX3.4), res_UMUX3.4) #homoscedascity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_UMUX3.4) #normal distribution check - dots should match line
```

```
qqline(res_UMUX3.4)
```

```
plot(density(res_UMUX3.4)) #normal distribution check - should have bell curve centred on  
0
```

```
#get output
```

```
#anova(lmm_UMUX3.4)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_UMUX3.4)
```

```
####F5
```

```
lmm_UMUX3.5 <- lmer(UMUX_Lite ~ BUS_5 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_UMUX3.5 <- residuals(lmm_UMUX3.5)
```

```

plot(fitted(lmm_UMUX3.5), res_UMUX3.5) #homoscedascity (equal variance) check
abline(0,0)

qqnorm(res_UMUX3.5) #normal distribution check - dots should match line
qqline(res_UMUX3.5)

plot(density(res_UMUX3.5)) #normal distribution check - should have bell curve centred on
0

#get output
#anova(lmm_UMUX3.5)
#info from anova are included in the summary() command
summary(lmm_UMUX3.5)

#####
###RELATION BETWEEN BUS AND NASA_TLX###
##Kendall's Tau BUS11
shapiro.test(new_data$BUS_score)
shapiro.test(new_data$NASA_workload)

cor.test(new_data$BUS_score, new_data$NASA_workload,use="pairwise.complete.obs",
method = "kendall")

#Plot of correlation
ggplot(new_data,aes(x=NASA_workload, y=BUS_score)) +

```

```
xlab("NASA_workload") + ylab("BUS11 overall scores")+  
geom_point() +  
geom_smooth(method = "lm",se = F, fullrange = F)  
  
##Kendall's Tau F1  
  
shapiro.test(new_data$BUS_1)  
  
shapiro.test(new_data$NASA_workload)  
  
cor.test(new_data$BUS_1, new_data$NASA_workload,use="pairwise.complete.obs", method  
= "kendall")  
  
#Plot of correlation  
  
ggplot(new_data,aes(x=NASA_workload, y=BUS_1)) +  
xlab("NASA_workload") + ylab("BUS factor 1")+  
geom_point() +  
geom_smooth(method = "lm",se = F, fullrange = F)  
  
##Kendall's Tau F2  
  
shapiro.test(new_data$BUS_2)  
  
shapiro.test(new_data$NASA_workload)  
  
cor.test(new_data$BUS_2, new_data$NASA_workload,use="pairwise.complete.obs", method  
= "kendall")  
  
#Plot of correlation
```

```
ggplot(new_data,aes(x=NASA_workload, y=BUS_2)) +  
  xlab("NASA_workload") + ylab("BUS factor 2")+  
  geom_point() +  
  geom_smooth(method = "lm",se = F, fullrange = F)  
  
##Kendall's Tau F3  
  
shapiro.test(new_data$BUS_3)  
  
shapiro.test(new_data$NASA_workload)  
  
cor.test(new_data$BUS_3, new_data$NASA_workload,use="pairwise.complete.obs", method  
= "kendall")  
  
#Plot of correlation  
  
ggplot(new_data,aes(x=NASA_workload, y=BUS_3)) +  
  xlab("NASA_workload") + ylab("BUS factor 3")+  
  geom_point() +  
  geom_smooth(method = "lm",se = F, fullrange = F)  
  
##Kendall's Tau F4  
  
shapiro.test(new_data$BUS_4)  
  
shapiro.test(new_data$NASA_workload)  
  
cor.test(new_data$BUS_4, new_data$NASA_workload,use="pairwise.complete.obs", method  
= "kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=NASA_workload, y=BUS_4)) +
  xlab("NASA_workload") + ylab("BUS factor 4")+
  geom_point() +
  geom_smooth(method = "lm",se = F, fullrange = F)
```

```
##Kendall's Tau F5
```

```
shapiro.test(new_data$BUS_5)
shapiro.test(new_data$NASA_workload)
```

```
cor.test(new_data$BUS_5, new_data$NASA_workload,use="pairwise.complete.obs", method
= "kendall")
```

```
#Plot of correlation
```

```
ggplot(new_data,aes(x=NASA_workload, y=BUS_5)) +
  xlab("NASA_workload") + ylab("BUS factor 5")+
  geom_point() +
  geom_smooth(method = "lm",se = F, fullrange = F)
```

```
##Linear Mixed Model
```

```
#checking linearity of the data
```

```
plot(x = new_data$BUS_score, y = new_data$NASA_workload)           #scatterplot
plot(x = new_data$BUS_1, y = new_data$NASA_workload)
plot(x = new_data$BUS_2, y = new_data$NASA_workload)
plot(x = new_data$BUS_3, y = new_data$NASA_workload)
```



```
plot(x = new_data$BUS_4, y = new_data$NASA_workload)
plot(x = new_data$BUS_5, y = new_data$NASA_workload)

#checking for clustering between NASA_workload and participants
boxplot(NASA_workload ~ participants, data = new_data, xlab = "participants", #boxplot
        ylab = "workload", main = "Clustering in workload scores by participants")

#checking for clustering between NASA_workload and chatbots
boxplot(NASA_workload ~ chatbot, data = new_data, xlab = "chatbot", #boxplot
        ylab = "workload", main = "Clustering in workload scores by chatbots")

##models for relationship with NASA_TLX - choose the one that is applicable
##1: random effect participants
#lmm_NASA <- lmer(NASA_workload ~ BUS_score + (1|participants), data = new_data)

#check distribution of residuals
#res_NASA1 <- residuals(lmm_NASA)
#plot(fitted(lmm_NASA), res_NASA1) #homoscedascity (equal variance) check
#abline(0,0)

#qqnorm(res_NASA1) #normal distribution check - dots should match line
#qqline(res_NASA1)

#plot(density(res_NASA1)) #normal distribution check - should have bell curve centred on 0
```

```
#get output

#anova(lmm_NASA)

#summary(lmm_NASA)

##2: random effect chatbot

#lmm_NASA2 <- lmer(NASA_workload ~ BUS_score + (1|chatbot), data = new_data)

#check distribution of residuals

#res_NASA2 <- residuals(lmm_NASA2)

#plot(fitted(lmm_NASA2), res_NASA2) #homoscedascity (equal variance) check

#abline(0,0)

#qqnorm(res_NASA2) #normal distribution check - dots should match line

#qqline(res_NASA2)

#plot(density(res_NASA2)) #normal distribution check - should have bell curve centred on 0

#get output

#anova(lmm_NASA2)

#summary(lmm_NASA2)

##3: random effect participants and chatbot

#BUS-11
```

```
lmm_NASA3 <- lmer(NASA_workload ~ BUS_score + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_NASA3 <- residuals(lmm_NASA3)
```

```
plot(fitted(lmm_NASA3), res_NASA3) #homoscedasticity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_NASA3) #normal distribution check - dots should match line
```

```
qqline(res_NASA3)
```

```
plot(density(res_NASA3)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_NASA3)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_NASA3)
```

```
#F1
```

```
lmm_NASA3.1 <- lmer(NASA_workload ~ BUS_1 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_NASA3.1 <- residuals(lmm_NASA3.1)
```

```
plot(fitted(lmm_NASA3.1), res_NASA3.1) #homoscedasticity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_NASA3.1) #normal distribution check - dots should match line
```

```
qqline(res_NASA3.1)
```

```
plot(density(res_NASA3.1)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_NASA3.1)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_NASA3.1)
```

```
#F2
```

```
lmm_NASA3.2 <- lmer(NASA_workload ~ BUS_2 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_NASA3.2 <- residuals(lmm_NASA3.2)
```

```
plot(fitted(lmm_NASA3.2), res_NASA3.2) #homoscedasticity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_NASA3.2) #normal distribution check - dots should match line
```

```
qqline(res_NASA3.2)
```

```
plot(density(res_NASA3.2)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_NASA3.2)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_NASA3.2)
```

```
#F3
```

```
lmm_NASA3.3 <- lmer(NASA_workload ~ BUS_3 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_NASA3.3 <- residuals(lmm_NASA3.3)
```

```
plot(fitted(lmm_NASA3.3), res_NASA3.3) #homoscedasticity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_NASA3.3) #normal distribution check - dots should match line
```

```
qqline(res_NASA3.3)
```

```
plot(density(res_NASA3.3)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_NASA3.3)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_NASA3.3)
```

```
#F4
```

```
lmm_NASA3.4 <- lmer(NASA_workload ~ BUS_4 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_NASA3.4 <- residuals(lmm_NASA3.4)
```

```
plot(fitted(lmm_NASA3.4), res_NASA3.4) #homoscedasticity (equal variance) check
```

```
abline(0,0)
```

```
qqnorm(res_NASA3.4) #normal distribution check - dots should match line
```

```
qqline(res_NASA3.4)
```

```
plot(density(res_NASA3.4)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_NASA3.4)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_NASA3.4)
```

```
#F5
```

```
lmm_NASA3.5 <- lmer(NASA_workload ~ BUS_5 + (1|participants) + (1|chatbot), data =  
new_data)
```

```
#check distribution of residuals
```

```
res_NASA3.5 <- residuals(lmm_NASA3.5)
```

```
plot(fitted(lmm_NASA3.5), res_NASA3.5) #homoscedasticity (equal variance) check
abline(0,0)
```

```
qqnorm(res_NASA3.5) #normal distribution check - dots should match line
qqline(res_NASA3.5)
```

```
plot(density(res_NASA3.5)) #normal distribution check - should have bell curve centred on 0
```

```
#get output
```

```
#anova(lmm_NASA3.5)
```

```
#info from anova are included in the summary() command
```

```
summary(lmm_NASA3.5)
```

```
#####Demographic data
```

```
#excluding participants who did not finish the scales for at least 1 chatbot
```

```
chatbot_experience <- c(chatbot_experience[1:40], chatbot_experience[42:50],
chatbot_experience[52:59], chatbot_experience[62:64], chatbot_experience[75],
chatbot_experience[80], chatbot_experience[83], chatbot_experience[87],
chatbot_experience[89:112], chatbot_experience[114:116], chatbot_experience[118:124],
chatbot_experience[126:136])
```

```
sex <- c(data$sex[1:40], data$sex[42:50], data$sex[52:59], data$sex[62:64], data$sex[75],
data$sex[80], data$sex[83], data$sex[87], data$sex[89:112], data$sex[114:116],
data$sex[118:124], data$sex[126:136])
```

```
age <- c(data$age[1:40], data$age[42:50], data$age[52:59], data$age[62:64], data$age[75],
data$age[80], data$age[83], data$age[87], data$age[89:112], data$age[114:116],
data$age[118:124], data$age[126:136])
```

```
eng <- c(data$English_proficiency[1:40], data$English_proficiency[42:50],
data$English_proficiency[52:59], data$English_proficiency[62:64],
data$English_proficiency[75], data$English_proficiency[80], data$English_proficiency[83],
data$English_proficiency[87], data$English_proficiency[89:112],
data$English_proficiency[114:116], data$English_proficiency[118:124],
data$English_proficiency[126:136])
```

```
#frequency tables to check for missing data
```

```
age %>% map(tabyl)
```

```
#getting demographics
```

```
summary(as.numeric(age))
```

```
sd(as.numeric(age))
```

```
data[c('sex', 'English_proficiency')] %>% map(tabyl)
```

```
data[c('Q67_1')] %>% map(tabyl)
```

```
chatbot_experience <- (as.numeric(data$Q82_1) + as.numeric(data$Q82_1) +
as.numeric(data$Q82_1)) / 3
```

```
summary(chatbot_experience)
```

```
sd(chatbot_experience)
```

```
#####Descriptives
```

```
summary(new_data$BUS_score)
```

```
sd(new_data$BUS_score)
```



```
summary(new_data$BUS_1)
```

```
sd(new_data$BUS_1)
```

```
summary(new_data$BUS_2)
```

```
sd(new_data$BUS_2)
```

```
summary(new_data$BUS_3)
```

```
sd(new_data$BUS_3)
```

```
summary(new_data$BUS_4)
```

```
sd(new_data$BUS_4)
```

```
summary(new_data$BUS_5)
```

```
sd(new_data$BUS_5)
```

```
summary(new_data$UMUX_Lite)
```

```
sd(new_data$UMUX_Lite)
```

```
summary(new_data$NASA_workload)
```

```
sd(new_data$NASA_workload)
```