Evaluating the Chatbot Usability Scale: A Psychometric and Designometric Perspective

Marlen Braun

24th of March 2023

1st Supervisor: Dr. Simone Borsci

2nd Supervisor: Dr. Martin Schmettow

Master's Thesis (25 EC)

University of Twente

BMS Faculty

Department of Psychology

Abstract

The rise of the world wide web and artificial intelligence led to an increase in conversational agents, especially chatbots. However, until recently, there was no singular instrument to assess user satisfaction with chatbots. This is why previously the Bot Usability Scale (BUS) was developed. Initially starting from 42 items, the most recent version is BUS-11 with 11 items. The present study was conducted to further validate the five-factor structure of the BUS-11, as well as its reliability and validity. Moreover, the BUS-11 was assessed from a psychometric and designometric perspective. Finally, other factors possibly affecting satisfaction with chatbots were examined.

A Structural Equation Model (SEM) was used to conduct a Confirmatory Factor Analysis (CFA) as well as examine relationships between satisfaction and cognitive workload, previous experience, and a disposition to trust. The psychometric sample consisted of 137 participants and the designometric sample consisted of 22 chatbots.

Results confirm the five-factor structure from a psychometric, but not from a designometric perspective. Moreover, the BUS-11 was shown to be a reliable and valid instrument. A negative relationship between satisfaction and cognitive workload was found.

This research further validated the BUS-11 from a designometric point of view and showed that it can be used to assess user satisfaction after chatbot interaction. Further research is needed into examining the factorial structure from a designometric perspective as the factorial structure is likely to be different from the psychometric perspective.

Keywords: Chatbots, Conversational Agents, User Satisfaction, Satisfaction, Interaction Satisfaction, Bot Usability Scale

1. Introduction

1.1. Conversational Agents and Chatbots

With the rise of the internet and artificial intelligence, internet users encounter more and more conversational agents in their life. A conversational agent is "software that accepts natural language as input" and uses this to communicate back and forth with the user (Griol et al., 2013, p. 760). This process is highly interactional as it goes beyond giving instructions to a computer (Allen et al., 2001). As the user interacts with a computer and not another human, this is described as computer-mediated communication (Hill et al., 2015). The complexity of conversational agents varies. For example, there are rule-based approaches in which answers are pre-defined, but also generative-based approaches in which the dialogue is formed more individually and based on training data (Adamopoulou & Moussiades, 2020; Hussain et al., 2019). Consequently, conversational agents differ in their ability to process complex input and provide individualized answers.

One commonly used type of conversational agent on the internet is chatbots. They are characterized by text-based communication that often does not go beyond small talk (Hill et al., 2015). Although the user is not interacting with another human, they typically do not encounter difficulties in understanding and being understood by the chatbot (Hill et al., 2015). Moreover, online interactions have become increasingly asynchronous and shorter in text. As chatbots sometimes use a similar interaction style, humans do not experience problems with this type of interaction (Dale, 2016). Hill et al. (2015) even found that users send up to double the number of text messages to chatbots compared to human assistants. It was not due to the chatbot's inability to comprehend but indicated that users are not hesitant in the interaction and not intimidated by the chatbot's potential power.

Chatbots have become increasingly prevalent in the commercial world. As the contact between consumers and companies is shifting more towards the virtual world and towards more "technology-dominant rather than human-driven" interaction, chatbots can create value (Larivière et al., 2017, p. 329). Brandtzaeg and Følstad (2017) researched different reasons why customers use chatbots. They found that most users reported "productivity" and "entertainment" to be the main motivators. Chatbots assist in finding information quickly and efficiently, therefore reducing the need to click through many different website pages (Brandtzaeg & Følstad, 2017; Følstad et al., 2018). Moreover, compared to a human they can handle multiple inquiries at the same time and are available at any hour of the day, making them more money- and time-efficient as well as quick to respond (Adamopoulou & Moussiades, 2020; Chung et al., 2020; Okuda & Shoda, 2018). As chatbots are not human, users might also feel less embarrassed to ask certain questions, thus decreasing the threshold

for interaction (Følstad et al., 2018). However, there are also negative sides: some chatbots cannot comprehend complicated questions which leads to confusing answers. Moreover, some users are concerned about the security of the entered information and if they can trust the chatbot with their entered private data (Følstad et al., 2018). Consequently, the design of a chatbot impacts customer satisfaction and customer experience.

1.2. Assessing User Satisfaction

Companies use customer satisfaction as a success metric. According to Bearden and Teel (1983), customer satisfaction impacts both complaint reports and customer attitude which in turn impacts customer intention. They found that fewer complaint reports are issued when satisfaction is high. Moreover, higher satisfaction leads to intentions like deciding to buy a product again or recommend it to friends. In a digital context, it was also found that satisfaction binds the customer to the product and ensures a pleasurable experience (Ashfaq et al., 2020; Chang & Chen, 2008; Chung et al., 2020). Additionally, the chatbot represents a company's brand, thus, a chatbot affects how a company is perceived (Youn & Jin, 2021). Here, the marketing of a product and the user experience are interrelated.

Assessing the satisfaction with chatbots specifically with one scale is challenging. In a systematic mapping study, Ren et al. (2019) found that the usability of chatbots is commonly assessed with questionnaires, interviews, or a combination of both. They found that usability included three different aspects: "effectiveness, efficiency and satisfaction" (Ren et al., 2019, p. 1682) . This is in line with the International Organisation for Standardization (ISO) standards which also list these three aspects as building stones of usability (International Organization for Standardization, 1998). Regarding satisfaction, they found that "ease of use" and "complexity control" were the most important factors when attempting to measure satisfaction (Ren et al., 2019, p. 1684) . Moreover, many questionnaires assess the direct communication of the chatbot with the user, for example, if it is able to keep track of context and ask questions accordingly (Ren et al., 2019). However, a common problem is that there is not one chatbot satisfaction scale that encompasses all aspects satisfactorily. For example, Chung et al. (2020) investigated the relationship between chatbot services of luxury brands and customer satisfaction and had to use a mixture of two previously constructed scales. One of these scales came from a study by Joosten et al. (2016) who also in turn built it from two other satisfaction scales (Matzler et al., 2005, as cited in Joosten et al., 2016; Oliver & Swan, 1989, as cited in Joosten et al., 2016). When different studies use different scales to assess chatbot satisfaction, the scales often assess different aspects of satisfaction. Therefore, the

results might not be comparable. Here, it becomes apparent that a general chatbot satisfaction scale is needed, especially as chatbots are becoming more and more popular.

To assess chatbot satisfaction, it is possible to use general technology satisfaction scales. One of the first and also commonly used scale is the System Usability Scale (SUS) which has ten items (Brooke, 1996). It was established to have a scale that could measure effectiveness, efficiency, and satisfaction as defined by the International Organization for Standardization (1998). However, there were some complications with this scale. With ten items, it is long and takes some time to complete. Moreover, some items posed difficulties for non-native English speakers. For example, item 8 "I found the system very cumbersome to use" was difficult to understand due to the word "cumbersome" (Finstad, 2010, p.323). Therefore, Finstad (2010) developed the Usability Metric for User Experience (UMUX) scale to have a scale that uses more commonly used words and reduces the number of items. The UMUX scale has four items, but in some usability testing situations, that is still long. Often, the UMUX scale is not used alone, but complementary to other scales that assess the product. As users then need to spend a lot of time and mental capacity to fill in all questionnaires, there was an incentive to create an even lighter version. This was achieved with the UMUX-LITE scale consisting of only two items (Lewis et al., 2013). It was found that there is a strong correlation between the results of all three scales, indicating that all of them can be used as a relatively quick way to assess satisfaction (Borsci et al., 2015).

These scales miss one aspect that is specific to interaction with chatbots. Chatbots are less static than other systems and incorporate a conversational aspect. This aspect is not covered by items in the SUS, UMUX, or UMUX-LITE. For example, the SUS includes items like "I thought the system was easy to use" or "I think that I would like to use the system frequently" (Brooke, 1996). These aspects apply to chatbots as well, but the quality of interaction or the conversation is not measured.

To overcome this, the Bot Usability Scale (BUS) was developed. Initially, this scale consisted of 42 items to capture all aspects that contribute to user satisfaction when interacting with chatbots (Borsci et al., 2022). As this is a long list of items and already the ten-item long SUS has been criticized for being too long, an effort was made to reduce the number of items. As an outcome, the BUS-15 was established which has five factors and a reliability between .76 and .87 (Borsci et al., 2021). Further, there is a strong correlation between the BUS-15 and the UMUX-LITE which indicates that it truly measures the construct of satisfaction. As a 15-item long scale is still long, a follow-up study investigated a reduced version called BUS-11 (see Appendix A). The five-factor structure was found with

these 11 items as well, however, more validation studies were recommended (Borsci et al., 2021).

1.3. Cognitive Workload

One of the factors that might influence satisfaction with a chatbot is cognitive workload. A high cognitive workload leads to "incomplete attention" and more human errors (Huey & Wickens, 1993, p. 85). This could be frustrating for the user and consequently lead to a lower level of satisfaction. Moreover, a high cognitive workload means that some information might be missed (Huey & Wickens, 1993). In an interaction with a chatbot, this could mean that information provided by the chatbot is not fully comprehended and the user gets frustrated because they cannot achieve their goal.

Cognitive workload influences perceived usability. It has been found that there is a negative relationship between perceived workload and perceived usability (Kokini et al., 2012). Similarly, it was found that cognitive workload and usability are two different constructs (Longo & Dondio, 2016). In a follow-up study, Longo (2018) found that cognitive workload and usability affect each other, and both also affect task performance. He suggested that the two measures can give a richer view of performance and complement each other. Further, it was found that when users had to interact with different bookstore websites and search for certain sections, the user satisfaction and cognitive load are significantly correlated (Schmutz et al., 2009). This research investigates further the relationship between satisfaction and cognitive workload. If workload affects satisfaction, then this can be used for design recommendations. For example, chatbots can be built and designed to induce a certain cognitive workload which leads to the highest satisfaction scores. Moreover, a high workload and thus, increased frustration might let the user abort the chatbot interaction.

1.4. Previous Experience

Previous experience might influence satisfaction as well. When measuring satisfaction with the SUS, it was found that higher satisfaction scores were found in users with more previous experience (McLellan et al., 2012). Similarly, Kortum and Johnson (2013) investigated how perceived usability scores change over time when users get trained and thus, gain experience with the system. In a group of users who got trained in Microsoft Publisher, they found that the perceived usability increased with experience. However, they did not find this effect in a second group who got trained in MathWorks MATHLAB. They mainly attributed this to an absence in experimental control that the Microsoft Publisher group had, as well as differences in the task (Kortum & Johnson, 2013). Therefore, previous experience

might still impact usability, and with that, satisfaction. Another study by Borsci et al. (2015) found that previous experience affected satisfaction scores as measured by the SUS, UMUX, and UMUX-LITE.

Investigating this relationship further can aid designers. For example, imagine a chatbot will be implemented in a system which is used by users who typically have less experience with chatbots. A high user satisfaction could then still be achieved by taking the level of experience into account and designing specifically for it. For example, features that are known to users with much experience might have to be explained first, or a general introduction to chatbots would have to be given. Therefore, this research investigates the relationship between previous experience and user satisfaction.

1.5. Disposition to Trust

Another factor possibly impacting user satisfaction is trust. It was found that nearly 95% of users have denied providing personal information to a website before, and around 40% have provided false data (Hoffman et al., 1999). Moreover, a study in Japan found that citizens would trust a chatbot less when it provides parenting advice compared to a chatbot which provides waste separation advice (Aoki, 2020). Here, it can be observed that when it comes to private, personal information, users tend to have less trust in chatbots. This might come from a fear of personal data being treated without high-security standards, especially in domains like finance (Følstad et al., 2018; Zamora, 2017).

To overcome mistrust, it is helpful to investigate what influences trust. Nordheim et al. (2019) found three main factors which impact trust. First, there are chatbot-related factors which for example include “perceived expertise and responsiveness” (Nordheim et al., 2019, p. 3). The competence aspect is especially important. The more competent, the more the user perceives the system to be useful and to be able to satisfy their demands and needs (Benbasat & Wang, 2005). Similarly, in a study which focused on operators’ interaction with automation, it was found that the operators trusted the machine more if it appeared to be competent (Muir & Moray, 1996). Competency is also a factor which is indirectly assessed in the BUS-11, for example, with item 6 “That chatbot gives me the appropriate amount of information” (Borsci et al., 2021). Second, trust is impacted by environmental-related factors, for example, perceived risk and how the specific brand is perceived (Nordheim et al., 2019). Thirdly, user-related factors, especially a disposition to trust, affect the level of trust when using a chatbot (Nordheim et al., 2019). Here, it has been found that a disposition to trust positively affects the trust in chatbots (Benke et al., 2022; Wang & Benbasat, 2008).

Disposition to trust affects trust behavior. McKnight et al. (2002) found that the disposition to trust affects trusting intentions which in turn, affect trust-related behavior. This trust-related behavior could include, for example, sharing sensitive information. This type of behavior is especially relevant when interacting with finance or healthcare chatbots. For example, if a healthcare chatbot has the task to assess which illness a user has, it is crucial for a correct assessment that that user enters the correct data, even if they are embarrassed about it. As chatbots are on the rise and there are users who have never or very little interacted with chatbots before, it is helpful to know how a disposition to trust impacts satisfaction. For example, if a chatbot commonly interacts with users who are known to have a low disposition to trust, the designers can intentionally include features to increase trust.

Disposition to trust is related to faith in humanity. Faith in humanity describes a tendency to have a trusting view of the world and to allow oneself to rely on others (McKnight et al., 2002). The disposition to trust is further split into four concepts: competence, benevolence, integrity, and trusting stance. An example statement for competence is “I believe that most professional people do a very good job at their work”, for benevolence it is “The typical person is sincerely concerned about the problem of others”, for integrity “In general, most folks keep their promise” and for trusting stance “I usually trust people until they give me a reason to not trust them” (McKnight et al., 2002).

This research investigates the effect of disposition to trust on user satisfaction and not which chatbot features lead to higher or lower trust. This is due to the scope and main focus on the validation of the BUS-11. To validate the BUS-11, study participants are required to answer several questionnaires after each interaction with a chatbot. Similarly, with the motivation to reduce the ten-item SUS scale to a four-item UMUX scale, the aim is to not overwhelm participants with too many questionnaires (Lewis et al., 2013). As researching the disposition to trust is relevant for the design of chatbots and as it requires participants to fill in only one trust questionnaire at the beginning of the research, it was decided to investigate this aspect of trust in this study.

1.6. Psychometric versus Designometric Perspective

In the field of Psychology, a commonly used perspective for analyzing data is the psychometric perspective. This means that individuals are compared to each other and ranked. Moreover, they are usually used to compare individual characteristics. A psychometric model includes at least humans and items. For example, a class of school children might get tested with an IQ test consisting of ten items. The collected data then contains the measure of each item for each school child. To compare the children, the IQ score is calculated by summing up

the response to the ten items per child and then averaging them. Then, the children can be compared and ranked according to their intelligence as measured by the IQ test. However, this is different when you want to compare designs instead of humans. For example, if you have a ten-item scale to assess usability and you only average the item responses per participant (and not per design), it means that you compare humans and not designs. This is a contradiction if your goal is to compare two designs, for example, to decide which design has a higher usability.

Therefore, Schmettow (2021) developed Designometrics. The goal is to compare and rank designs instead of persons. Here, the "measurement is the encounter of three populations, humans, items, and ultimately, designs" (Schmettow, 2021, p. 261). These three measures form a cuboid. To get a psychometric dataset, the responses are averaged across items to create a person-level score. In contrast, to get a designometric dataset, the responses need to be averaged over the designs. These different designs could for example be different chatbots. When gathering the data for a cuboid, it means that every participant rates the items after using or interacting with different designs (Schmettow, 2021). If the responses were averaged over the designs, a designmetric matrix is achieved. Then, standard psychometric procedures can be applied to this dataset.

Scales are commonly developed using the psychometric approach. However, when scales are developed to compare designs and not humans, they should be assessed from a designometric and not a psychometric viewpoint. In a designometric study, the population consists of designs instead of humans. Therefore, a large number of designs are needed. In practice, this is difficult because the population of designs needs to be big enough to estimate complex models like a Structural Equation Model.

In this research, both psychometric and designometric models are used. The BUS-11 was developed from a psychometric viewpoint. The reason was that there was no other way to develop the scale in the past. Therefore, this study assesses the BUS-11 from a psychometric perspective for comparison reasons. Additionally, this study takes the designometric perspective. The reason is that the purpose of BUS-11 is to compare designs and not humans.

The main focus of this research is to further validate BUS-11. This study investigates how reliably BUS-11 measures satisfaction and how well it can distinguish between different designs. Under a designometric model, it estimates how well a chatbot satisfies users. Under a psychometric model, this estimates how satisfied a particular person is with chatbots. This can be helpful when assessing how satisfied different groups of people are with a chatbot. For example, there might be a difference between people who have lots of experience using digital interfaces versus people who have little experience. If people with little experience are

less satisfied, this has implications for the design process. For example, if you want to design a chatbot and you know that the majority of your users will have little experience, then you need to design differently compared to when your users would have lots of experience. In these cases, the psychometric perspective is valuable.

As the BUS-11 was first developed from a psychometric perspective and has the aim to compare designs and not humans, this research examines the factorial structure of the BUS-11 from a psychometric and designometric perspective. Moreover, it compares the factorial structure of both perspectives with each other.

1.7. The Current Study

The main aim of the present research is to validate the five-factor structure of BUS-11. The first research question concerns this factorial structure. It has been found previously by Borsci et al. (2021), however, further validation was needed. Therefore, the first research question is "*Is the construct of the BUS-11 composed of five factors, in line with previous research by Borsci et al. (2021), when tested from the psychometric and designometric perspective?"* (RQ1).

To be used on a larger scale, a scale must be reliable. Therefore, the second research question is "*How reliable are the items of BUS-11?*" (RQ2).

Moreover, Borsci et al. (2021) found a strong correlation between the BUS-11 and the UMUX-LITE, indicating that both measure satisfaction. This research aims to investigate the concurrent validity of the BUS-11 with the UMUX-LITE to determine if the BUS-11 still measures satisfaction satisfactorily. Therefore, the third research question is "*Are the BUS-11 and UMUX-LITE correlated (concurrent validity)?"* (RQ3).

As discussed above, different individual factors might affect satisfaction. In this research the following factors are considered:

1) Cognitive workload. It was found that perceived usability and cognitive workload are negatively related (Kokini et al., 2012).
2) Previous experience. Higher satisfaction scores measured by the SUS are related to a previous experience (McLellan et al., 2012) and a previous experience affects satisfaction measured by the SUS, UMUX, and UMUX-LITE (Borsci et al., 2015).
3) A disposition to trust. Trust determines which information is shared with a chatbot, especially when it comes to personal data (Aoki, 2020; Zamora, 2017). Disposition to trust is on of the determinants of trust (Nordheim et al., 2019) and is positively related to it (Benke et al., 2022; Wang & Benbasat,

2008). The level of the disposition to trust might guide chatbot design decisions.

Therefore, the fourth research question is *"What is the relationship between the three individual factors of cognitive workload, previous experience and a disposition to trust, and satisfaction with chatbots measured by the BUS-11?"* (RQ4).

2. Methods

2.1. Sample

A convenience sample of 137 participants consisting of students from the University of Twente was taken. The participants received information about the content of the study and then consented to participate in it. They were informed that they could withdraw from the study at any point. The study was approved by the Ethics Committee of the Faculty of Behavioral, Management and Social Sciences (BMS) of the University of Twente. The participants received course credits in exchange for participating in the study via the BMS Test Subject Pool Website (University of Twente, 2022).

Under the psychometric perspective, the population does not consist of humans, but of designs (Schmettow, 2021). Therefore, the chatbot population used in this study is listed in this section. Data from ten different chatbots was collected (see Table 1). To increase the sample size, data by Huijsmans (2022) was added to the existing designometric dataset of this study. This included 52 additional participants who rated twelve additional chatbots.

Table 1

Chatbots used for the data collection of this study

Company name	Link to the chatbot
Adobe	https://helpx.adobe.com/support.html
Figma	https://help.figma.com/hc/en-us
Dropbox	https://www.dropbox.com/support
Access Bank	https://www.accessbankplc.com/pages/customer-support.aspx
Air New Zealand	https://www.airnewzealand.co.nz/help-and-contact
Royal Bank of Scotland	https://www.rbs.co.uk/support-centre.html#cora
Singapore Airlines	https://www.singaporeair.com/en_UK/sg/support/kris-the-chatbot/
Buoy Health Checker	https://www.buoyhealth.com/symptom-checker/
Benefit cosmetics	https://www.benefitcosmetics.com/en-us/contact-us
Singapore Government	https://www.gov.sg/

2.2. Measures

To measure satisfaction, the BUS-11 was used (Borsci et al., 2021). It is based on the BUS-15 which has been found to have an estimated reliability between 0.76 and 0.87 (Borsci et al., 2022). Moreover, when assessing the satisfaction of ten chatbots with the BUS-15 and the UMUX-LITE, a correlation of 0.61 to 0.817 was found between the BUS-15 and UMUX-LITE, thereby showing a good validity of the BUS-15 as well (Borsci et al., 2022). The BUS-11 items are rated on a 5-point Likert Scale ranging from 0 (Strongly disagree) to 5 (Strongly agree).

Next to the BUS-11, this study also used the UMUX-LITE to assess the concurrent validity of the BUS-11. Consisting of two items, it is short and saves time during administration. Moreover, it has been used in previous studies to assess the validity of the BUS-11 (Borsci et al., 2015). It has been shown to have a high reliability of 0.81 and 0.87, as well as a high concurrent validity of 0.81 with the SUS (Lewis et al., 2013). The UMUX-LITE items are rated on a seven-point Likert scale ranging from 0 (Strongly disagree) to 7 (Strongly Agree).

To measure cognitive workload, the Rating Scale Mental Effort (RSME) was used (Zijlstra & Doorn, 1985). It has a test-retest reliability of 0.78 and a moderate correlation of 0.55 with the “Schaal Ervaren Belasting” (SEB), a scale which measures experienced cognitive load (Zijlstra & Doorn, 1985). The scale consists of one item and is measured on a scale from zero to 150. Due to the shortness, it was chosen over other commonly used scales

like the NASA Task Load Index which consists of six items (Hart & Staveland, 1988). Thereby, the overall number of questions that the participants had to answer was reduced.

Although a one-item scale has limitations like it not being possible to measure its reliability, using it to assess cognitive workload was deemed acceptable. For every one of the five chatbots, each participant had to fill in the BUS-11, UMUX-LITE, as well as a measure of cognitive workload. It was chosen to keep the scale to assess workload short to keep the motivation of the participants and prevent them from dropping out of the study.

Previous experience was measured using the item "Please indicate how often you use chatbots". It was assessed on a five-point Likert scale ranging from 0 (Never) to 5 (Daily).

To measure a disposition to trust, subscales developed by McKnight et al. (2002) were used. Their main scale measures four constructs: disposition to trust, institution-based trust, trusting beliefs, and trusting intentions. This study used the scales for disposition to trust which are further divided into the four subconstructs of benevolence, integrity, competence, and trusting stance (McKnight et al., 2002). The reliability of the subconstruct items is high and Cronbach's alpha varied from 0.82 to 0.9. The concurrent validity of the four subconstructs with lambda coefficients ranged from 0.71 to 0.92 (McKnight et al., 2002). The items are rated on a 5-point Likert scale ranging from 0 (Strongly disagree) to 5 (Strongly agree).

2.3. Task

There was one task for each chatbot. It was written in English and the participant needed to use the chatbot to complete the task (see Appendix B). The task sometimes included background information regarding what the product was about or regarding a context. For example, "Figma is a digital prototyping and brainstorming tool" or "You will fly directly from Amsterdam to Singapore in three days" (see Appendix B). Thereby, the participants received sufficient information to complete the task without having to search for information about the product. The participant had to search for the chatbot on the website and then complete the task through interaction with the chatbot.

2.4. Procedure

Participants accessed the link to the study via the University of Twente Sona-Systems platform, a platform that is used by researchers to gather participants (University of Twente, 2022). Then they were directed to the online platform Qualtrics where the data was collected (Qualtrics, 2022). They were presented with information about the study and a consent form. If the participant did not consent, the study was stopped. If the participant consented, they

were asked to fill in the information regarding their demographic data (age, gender, nationality) as well as about familiarity with chatbots and disposition to trust. After that, the participant was randomly shown one of the ten chatbots.

For each chatbot, the participant was first shown the task and the link to the chatbot. They were instructed to open the link in an incognito mode in their browser as sometimes some chatbots did not appear otherwise. Once the participant found the chatbot, they interacted with it until they completed the task or decided that it was not possible to complete it. The participant then went back to the browser tab in which they had the survey open. Here, they first were asked if they were able to finish the task. Then they filled in the RSME, BUS 11, and UMUX-LITE scales. With that, they have completed all actions for this chatbot.

After that, they were randomly redirected to the next chatbot. In total, each participant interacted with five chatbots to be able to conduct a designometric analysis (Schmettow, 2021). After the fifth chatbot, the participant got the message that the study was finished, and they were thanked for their participation.

2.5. Data Analysis

The data was cleaned and analyzed by using Excel and R (Microsoft, 2022; R Core Team, 2022). The data was downloaded from Qualtrics. Due to the data format given by Qualtrics, the data was rearranged in Excel and unnecessary empty cells were deleted. 37 observations were excluded as the participant could not find the chatbot, for example, by indicating “could not find chatbot” in the question of whether or not they could complete the task. Then, the data was exported to R for further analysis.

For the factor analysis, two different datasets were created: one for the psychometric perspective and one for the designometric analysis. From the designometric perspective, the responses were averaged per chatbot to create an item-by-design response matrix. This reduced the cuboid of participants, chatbot, and items to a chatbot and item matrix (Schmettow, 2021).

To answer the first research question, a confirmatory factor analysis (CFA) was conducted. This analysis was conducted once with a designometric dataset and once with the psychometric dataset. The R package “lavaan” was used for this analysis (Rosseel, 2012). Previous validation studies, for example, by (Borsci et al., 2021), to assess the factor structure of the BUS-11 used the following criteria by Hu and Bentler (1999) to determine if the model is acceptable:

- Tucker-Lewis Index (TLI) ≥ 0.95
- Comparative Fit Index (CFI) ≥ 0.95
- Standardized Root Mean Square Residual (SRMR) ≤ 0.08
- Root Mean Square Error of Approximation (RMSEA) ≤ 0.06

Previously, the factorial structure of the BUS-11 was evaluated based on these criteria. As the goal of this research is to further validate the factorial structure, the same criteria are used again for the data collected in this study. That makes it possible to compare the results.

To answer the second research question about the reliability of the BUS-11, Cronbach's alpha was calculated. The criteria for good reliability was set to ≥ 0.7 (Cortina, 1993; Taber, 2018).

To answer the remaining research questions, the scores of the BUS-11, UMUX-LITE, RSME, disposition to trust, and previous experience were re-scaled to a scale of 0 – 1 using the R package "bayr" (Schmettow, 2022). This way all responses were scaled to a uniform interval. The existing CFA model was extended to a Structural Equation Model (SEM) by adding regressions. To answer the third research question about scale validity, a regression between the UMUX-LITE and BUS-11 scores was calculated. To answer the fourth research question regarding the relationship between cognitive workload, disposition to trust, and previous experience with satisfaction, regressions were used.

The remaining relationships between cognitive workload, disposition to trust, and previous experience with satisfaction were assessed using regressions as part of the SEM. The relationship was assessed by looking at the parameter estimate, as well as the upper and lower bounds of the confidence interval. As cognitive workload, disposition to trust and previous experience concern individuals and do not compare designs, only a psychometric model was used for this part.

3. Results

3.1. Descriptive Statistics

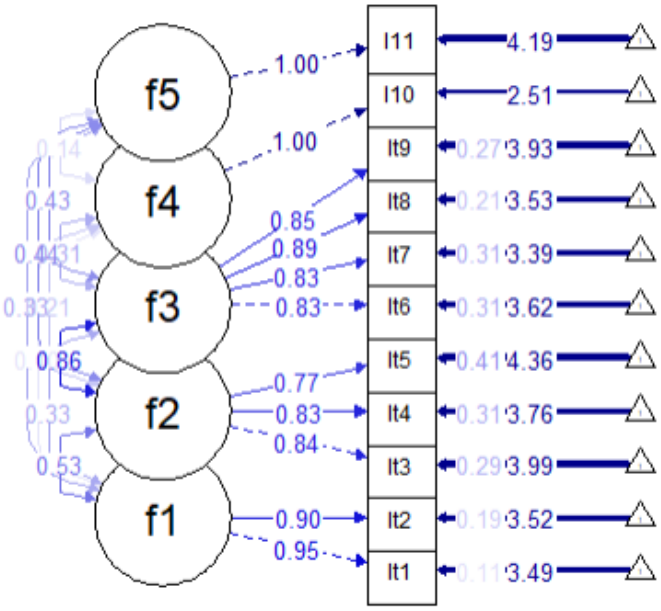
The sample consisted of 137 participants (M_{age} = 20.71, SD_{age} = 2.58, 92 female). About 27.74 % of the participants were Dutch, 48.18% were German, and 24.09 % had another nationality. On average, they reported being familiar with chatbots ($M_{familiarity}$ = 3.47, $SD_{familiarity}$ = 1.08), knew how chatbots work ($M_{knowledge}$ = 3.27, $SD_{knowledge}$ = 1.12), and felt confident to use a chatbot ($M_{confidence}$ = 3.36, $SD_{confidence}$ = 1.02). Regarding how often they use chatbots, 20.44% reported never using chatbots, 73% reported using them seldomly, and 7.3% reported using them more than one time per week.

3.2. Confirmatory Factor Analysis: Psychometric perspective

A confirmatory factor analysis was performed to test the five-factor structure of the BUS-11 from a psychometric perspective. The Comparative Fit Index and Tucker-Lewis Index are higher than .95 and thus indicate a good fit (CFI = .978, TLI = .966). The Standardized Root Mean Square Residual is lower than .08, indicating a good fit (SRMR = .26). The Root Mean Square Error of Approximation is higher than .06, thus indicating an insufficient fit (RMSEA = .066). The factor loadings range from 0.77 to 1. For factors consisting of more than one item, the factor loadings vary, but are smaller than 1. For example, for factor 3 the factor loadings of item 1 and item 2 are 0.95 and 0.95 respectively. In contrast, factor loadings for factor 2 are slightly lower. Here, the factor loadings for items 3, 4, and 5 are 0.84, 0.83, and 0.77 respectively. Generally, all items explain at least 77% of the variance in each factor (see Figure 1).

Figure 1

Visualization of the BUS-11 factor structure from a psychometric perspective



The correlations between the factors range from 0.096 to 0.863 (see Table 2). The strongest correlation can be observed between factors 2 and 3. Moreover, factors 1 and 2 are moderately correlated.

Table 2

Correlation between the factors of the BUS-11 under the psychometric model

Factor	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1					
Factor 2	.526**				
Factor 3	.328**	.863**			
Factor 4	.096*	.214**	.312**		
Factor 5	.330**	.441**	.429**	.136**	

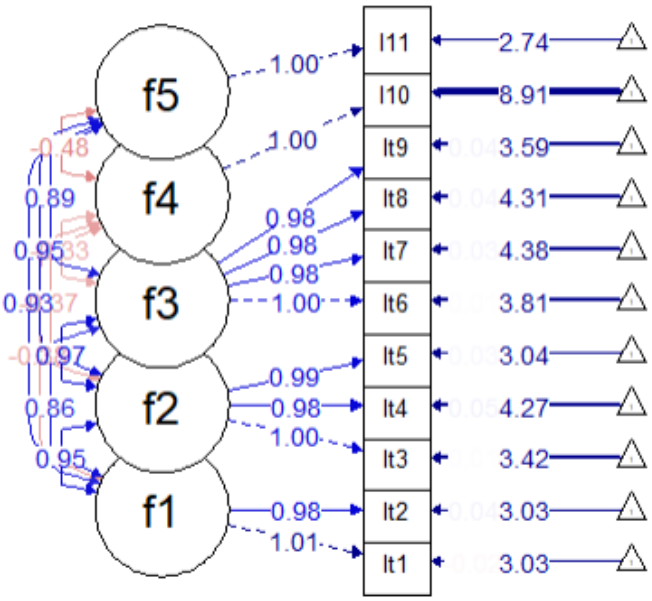
Note. *p < .05. **p < .01.

3.3. Confirmatory Factor Analysis: Designometric perspective

A confirmatory factor analysis was performed to test the five-factor structure of the BUS-11 from a designometric perspective. The Comparative Fit Index and Tucker-Lewis Index are lower than .95 and thus indicate a poor fit based on these criteria (CFI = .885, TLI = .85). The Standardized Root Mean Square Residual is lower than .08, indicating a good fit (SRMR = .024). The Root Mean Square Error of Approximation is higher than .06, thus indicating an insufficient fit (RMSEA = .275). The factor loadings range from 0.98 to 1.01. For every factor, there is at least one item that has a factor loading of 1. Moreover, all items explain at least 98% of the variance in each factor (see Figure 2).

Figure 2

Visualization of the BUS-11 factor structure from a designometric perspective



The correlations between the factors range from -0.381 to 0.049 (see Table 3). Correlations between factors 1, 2, and 3 are strong. Moreover, factors 1, 2 and 3 are strongly correlated with factor 5. Although not significant, factor 4 is negatively correlated with factors 1,2 and 3.

Table 3

Correlation between the factors of the BUS-11 under the designometric model

Factor	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1					
Factor 2	.949**				
Factor 3	.862**	.965**			
Factor 4	-.381	-.370	-.334		
Factor 5	.930**	.949**	.889**	.477*	

Note. *p < .05. **p < .01.

When using the outcome criteria to compare the psychometric perspective with a sample of 137 participants and the designometric perspective with a sample of 22 chatbots, the psychometric perspective indicates a better fit of the previously found five-factor structure of the BUS-11 (see Table 4).

Table 4

Outcome criteria of the confirmatory factor analysis of the BUS-11 using a psychometric dataset with 137 participants and designometric dataset with 22 participants

Perspective	CFI	TLI	SRMR	RMSEA
Psychometric	.978	.966	.026	.066
Designometric	.885	.850	.024	.275

3.4. Reliability

The internal consistency of BUS-11 as measured by Cronbach's alpha is high from the psychometric perspective (α = .89) and the designometric perspective (α = .93).

An inter-item correlation matrix was produced for both the psychometric and designometric datasets. In the psychometric dataset, the correlations varied from 0.094 (item 1 and item 10) to 0.852 (item 1 and item 2). In the designometric dataset, the correlations varied from 0.082 (item 10 and item 11) to 0.981 (item 1 and item 2). In the psychometric dataset, the correlations of items that belong to the same factor are stronger compared to correlations

of items that do not belong to the same factor. For example, scores of items 1 and 2 are strongly correlated, whereas scores of items 1 and 10 are weakly correlated (see Table 5).

In the psychometric dataset, item scores belonging to the same factor tend to be strongly correlated. Contrastingly, in the designometric dataset some item scores which do not belong to the same factor are highly correlated. For example, the scores of item 5 which belongs to factor two are strongly correlated with the scores of items 6 and 7 which both belong to factor 3 (see Table 6).

Table 5

Correlation of the BUS-11 from a psychometric perspective

Item	1	2	3	4	5	6	7	8	9	10	11
1											
2	.852**										
3	.441**	.438**									
4	.352**	.327**	.718**								
5	.439**	.392**	.666**	.598**							
6	.290**	.269**	.622**	.686**	.605**						
7	.227**	.206**	.566**	.590**	.578**	.661**					
8	.248**	.245**	.593**	.651**	.563**	.732**	.758**				
9	.300**	.286**	.588**	.653**	.546**	.702**	.711**	.766**			
10	.094**	.081*	.164**	.203**	.157**	.287**	.368**	.257**	267**		
11	.308**	.304**	.340**	.343**	.375**	.329**	.391**	.370**	.373**	.136**	

Note. *p < .05. **p < .01.

Table 6

Correlation of the BUS-11 from a designometric perspective

Item	1	2	3	4	5	6	7	8	9	10	11
1											
2	.981**										
3	.752*	.747*									
4	.626	.587	.940**								
5	.500	.501	.864*	.857**							
6	.525	.530	.916**	.914**	.850**						
7	.136	.150	.678*	.718*	.757*	.833**					
8	.369	.390	.751*	.720*	.607	.912**	.890**				
9	.409	.413	.765**	.800**	.639*	.916**	.882**	.956**			
10	.431	.420	.372	.351	.549	.450	.407	.364	356		
11	.485	.541	.450	.320	.209	.432	.376	.531	.536	.082	

Note. *p < .05. **p < .01.

3.5. Concurrent Validity

A regression as part of the SEM from the psychometric perspective was conducted between the scores of the BUS-11 and UMUX-LITE. It showed that BUS-11 and UMUX-LITE scores in this dataset were positively related (Estimate = 0.42, 95% CI [0.33, 0.51]) (see Figure 3). This indicates that BUS-11 and UMUX-LITE measure the same construct. Out of all five factors of the BUS-11, only factor 3 was significantly related with the UMUX-LITE scores (see Table 7).

Figure 3

Visualization of the relationship between BUS-11 and UMUX-LITE scores

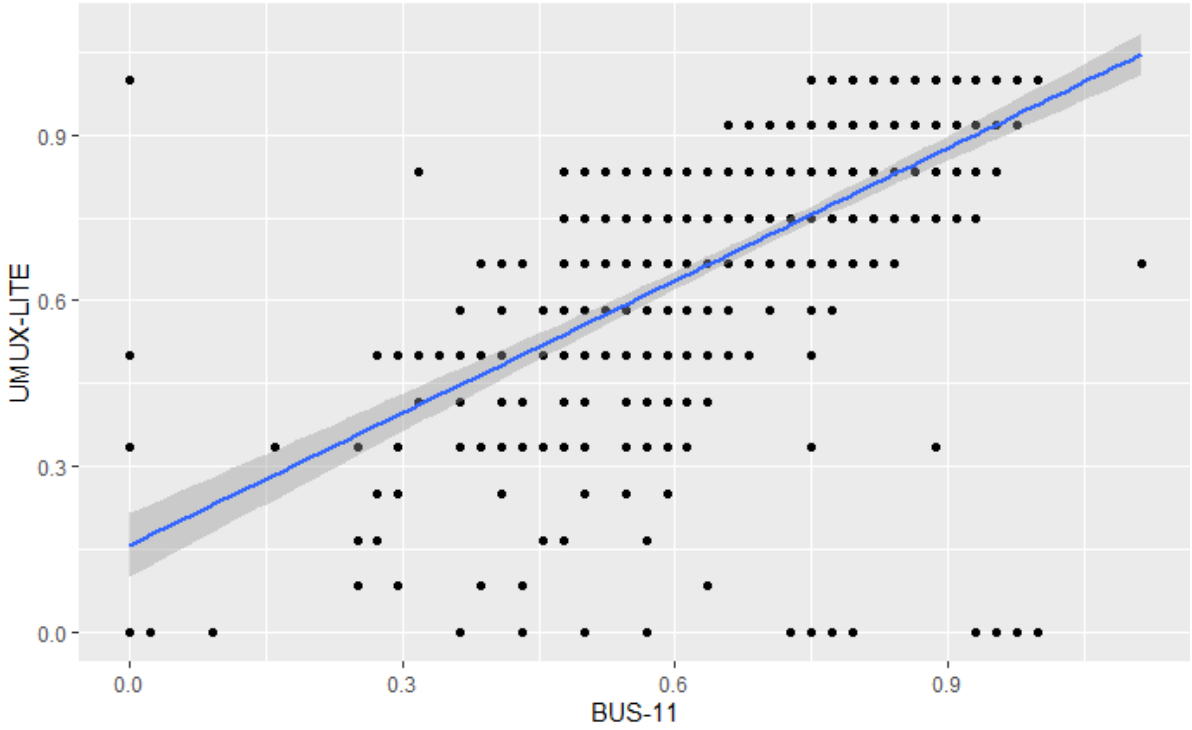


Table 7

Estimates of the regression between the five factors of the BUS-11 and UMUX-LITE scores

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
UMUX-LITE	.014	.034	.144**	-.017	.015

Note. *p < .05. **p < .01.

3.6. Other Factors affecting Satisfaction

Regressions of the cognitive workload, experience, and disposition to trust scores were conducted on the BUS-11 score as part of the SEM (see Table 8). A negative relationship between workload and BUS-11 scores was found (Estimate = -0.13, 95% CI [-0.19, -0.07]) (see Figure 4). Although the upper bound of the confidence interval is close to 0, this still indicates that there is a slight negative relationship between the two variables. None of the factors of the BUS-11 were significantly related to cognitive workload. A positive relationship was found between previous experience and BUS-11 (Estimate = 0.49, 95% CI [--1.54, 2.51]) (see Table 8). However, the confidence interval includes the value zero. Therefore, it cannot be certainly concluded that there is a positive relationship between previous experience and the BUS-11 score. Finally, a positive relationship was found between the disposition to trust and the BUS-11 score (Estimate = 0.04, 95% CI [-0.05, 0.13]). However, the confidence interval also includes the value zero. Therefore, it cannot be certainly concluded that there is a positive relationship between the disposition to trust and the BUS-11 score.

Table 8

Estimates and upper and lower bounds of the confidence intervals (CI) of the regressions between additional factors which might influence satisfaction

	Estimate	CI lower bound	CI upper bound
Cognitive workload	-.13	-.19	-.07
Previous experience	.49	-1.54	2.508
Disposition to trust	.04	-.05	.13

Figure 4

Visualization of the relationship between BUS-11 and RSME scores

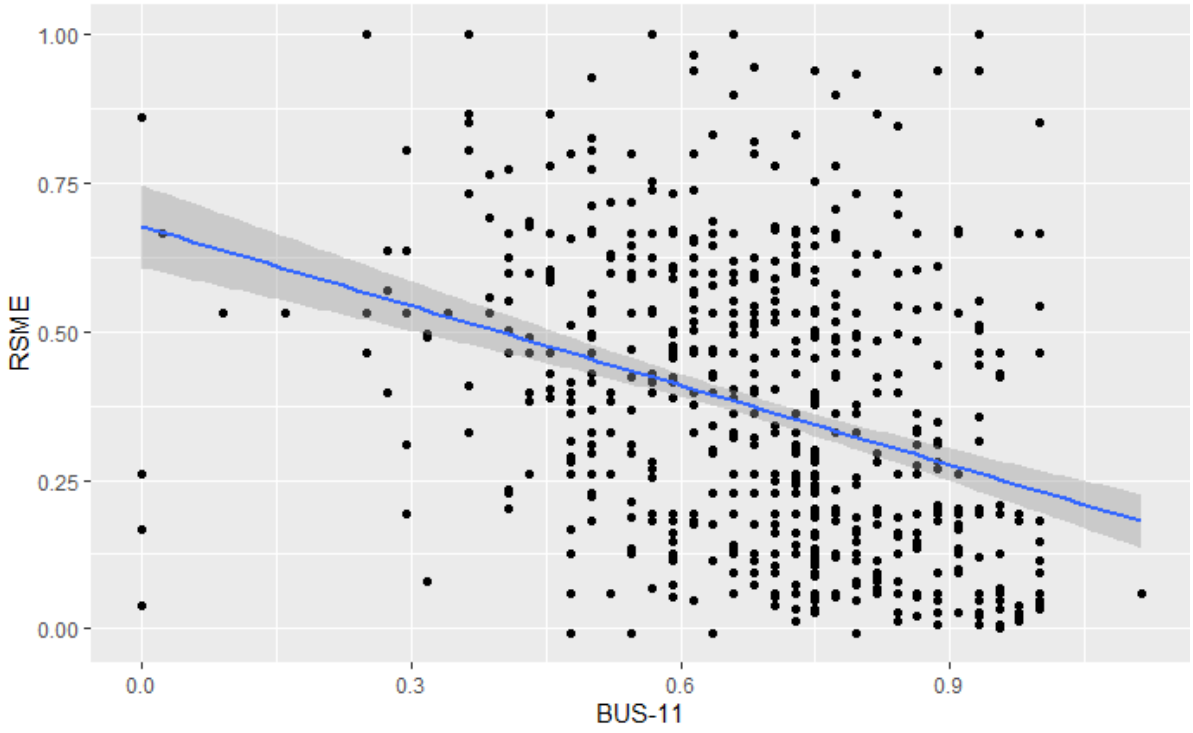


Table 9

Estimates of the regression between the five factors of the BUS-11 and RSME

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
RSME	-.001	-.029	-.059	.014	-.011

Note. *p < .05. **p < .01.

4. Discussion

The main purpose of this study was to further validate the factorial structure of the BUS-11. A CFA was conducted with a psychometric and designometric dataset. The five-factor structure was confirmed from a psychometric perspective, but not from a designometric perspective. Further, BUS-11 was confirmed to be a reliable and valid measurement instrument. Finally, a negative relationship between satisfaction measured by the BUS-11 and cognitive workload was found.

4.1. Factorial Structure

The first research question was “Is the construct of the BUS-11 composed of five factors, in line with previous research by Borsci et al. (2021), when tested from the psychometric and designometric perspective?”. The results suggest that from a psychometric perspective, the factorial structure is in line with previous research. The five-factor structure was shown to be a good fit. Moreover, the factor loadings are high and correlations between factors are moderate. The strongest factor correlations were found between factor 2 and factor 3. This might be because although the factors are different, they measure similar constructs. Factor 2 measures the “Perceived quality of the chatbot”, whereas factor 3 measures the “Perceived quality of conversation and information provided” (see Appendix A). For example, compared to factor 4 “Perceived privacy and security”, they are more similar to each other. Generally, the findings suggest that from a psychometric model, there are five different factors.

The results of this research did not support the previously found five-factor structure from a designometric perspective. First, factor loadings are high, but the correlations between factors are strong. A low to moderate correlation between factors could be expected as they all measure variables related to the theme of user satisfaction. However, very strong correlations between factors suggest that they measure the same construct and could be grouped into common factors. Therefore, a model with less than five factors might be more fitting from the designometric perspective. As all five factors are strongly correlated with each other, this points towards a one-factor model.

Second, the model fit of the designometric model is worse than that of the psychometric model. This might be because the factorial structure from a designometric perspective is generally different from this perspective. Moreover, one criterion that differed greatly was the RMSEA. In the designometric model, it was a lot higher. According to Kenny et al. (2015), the RSMEA has difficulties with models that have less complexity and a few degrees of freedom. Thus, the high RSMEA might also be explained by the small sample size in the designometric model and might improve if the sample size is bigger. According to Schmettow (2021), there are complications when using these criteria. Although a smaller sample size of chatbots can lead to these results, it might also be due to the fact that the factorial structure of the BUS-11 simply differs between the psychometric and the designmetric perspective.

4.2. Reliability and Validity

The second research question was “How reliable are the items of BUS-11?”. Generally, the internal consistency of the BUS-11 from both the psychometric and designometric perspectives was high. Therefore, it can be concluded that the scores of BUS-11 are reliable. Further, item scores belonging to the same factor were more strongly correlated compared to those not belonging to the same factor in the psychometric dataset. This is because they measure the same latent variable and are also more conceptually related. However, for the designometric dataset, also item scores not belonging to the same factor are strongly correlated. This might be because the factorial structure from the designometric perspective is generally different from the five-factor structure from the psychometric perspective. It is a similar trend to the strongly correlated factors in the designometric model. There, a strong correlation between factors indicates that they might measure similar latent variables. Therefore, it makes sense that also many items, including those not belonging to the same factor, are strongly correlated.

The third research question was “Are the BUS-11 and UMUX-LITE correlated (concurrent validity)?”. It was found that there is a strong positive correlation between BUS-11 and UMUX-LITE. This means that they measure a similar construct which is in line with previous research (Borsci et al., 2022). This means that BUS-11 is a valid instrument to measure user satisfaction. This solves the previously identified gap in available instruments to measure elements that are specific to chatbot interaction, for example, the conversational aspect. Other instruments currently available do not encompass all aspects satisfactorily and researchers often had to use mixtures of different scales (Chung et al., 2020; Joosten et al., 2016). With the BUS-11, this problem is solved. This is especially relevant in times when chatbots and artificial intelligence (AI) are becoming more and more important in the digital world.

4.3. Other Factors affecting Satisfaction

The fourth research question was “What is the relationship between the three individual factors of cognitive workload, previous experience, and a disposition to trust, and satisfaction with chatbots measured by the BUS-11?”.

A negative relationship between satisfaction measured by the BUS-11 and cognitive workload was found. This means that there was higher satisfaction in interactions where the workload was low. This is in line with previous research (Kokini et al., 2012; Schmutz et al., 2009). This means that if a designer wants to create a chatbot that users are satisfied with, they have to consider the cognitive workload. In their design process, they need to make sure

that the workload remains low for the user. For example, this could be achieved by having a simple screen design, chunking information, and avoiding that the user must multi-task (Grunwald & Corsbie-Massay, 2006).

No relationship was found between satisfaction measured by the BUS-11 and previous experience. This is not in line with previous research (Borsci et al., 2015; McLellan et al., 2012). This might be due to using an unstandardized way of measuring previous experience. Moreover, the sample contained mainly university students. This population tends to be more experienced with technology and the world wide web in general. Thus, while they might not interact with chatbots a lot, they might have a technology affinity. However, as this study measured the previous experience with chatbots and not technology affinity or competence in general, this might have led to non significant results.

No relationship was found between satisfaction measured by the BUS-11 and a disposition to trust. This might be due to several factors. First, in this research, the decision was made to research the disposition to trust and not other trust-related factors. This was because the literature also suggested that a disposition to trust might affect satisfaction (McKnight et al., 2002; Nordheim et al., 2019). Moreover, participants already had to spend much time responding to several scales after every chatbot interaction. Adding several trust-related items to this would have made the participants even more tired and might have resulted in dropouts.

While a disposition to trust was not related to satisfaction in this study, there might be another relationship between trust and satisfaction. One angle could be that there might be certain features which make the chatbot more trustworthy. It was found that visual and textual features of the chatbot can affect how trustworthy a chatbot is perceived (Følstad et al., 2018; Toma, 2010). This concerns especially how human-like the chatbot is (Følstad et al., 2018). Additionally, Przegalinska et al. (2019) suggest three new dimensions of trust, for example, honest and transparent communication. More research into these aspects is needed. Researching trust in chatbots remains crucial because of its relevance for new users to start using chatbots and for established users to continue using them (Corritore et al., 2003).

4.4. Limitations

There are two main limitations to this study. First, the sample size for the designometric model was small. This means that the results must be viewed with caution. Creating a large sample for a designometric study is difficult. It requires a bigger time commitment from human participants as they have to interact with more chatbots. In this study, additional data from Huijsmans (2022) was used. The data collected for this study

included ten chatbots, and an additional sample of 12 chatbots from a previous study was added. Letting one participant interact with and rate more than five to six chatbots will increase tiredness and potentially lead to more dropouts. Therefore, the choice was made to not include more chatbots and instead, use the data collected in this study for potential future studies.

Second, the previous experience was measured using an unstandardized way. This means that the results need to be viewed with caution.

4.5. Future Research

4.5.1. Factorial Structure from the Psychometric and Designometric Perspective

In this research, the five-factor structure of the BUS-11 was confirmed from a psychometric, but not a designometric perspective. The results suggest that the factorial structure of the BUS-11 is different from a designometric perspective, for example, it could be a one-factor structure. To find out what the factorial structure is from the designometric view, future research must focus on collecting data for more chatbots and combining it with previous data so that future designometric models can be more accurate.

Moreover, going back to the original 42-item BUS and performing an exploratory factor analysis (EFA) instead of a CFA would be ideal. As the reduction from 42 to 11 items was done from a psychometric view, this might be different from a designometric point of view. For example, the types and numbers of items the BUS could be reduced to from the original 42 items might not be the same as the current BUS-11. However, the same problem of data availability applies here. Collecting the data for the BUS-11 is already a huge time commitment from the participant side. Doing this again for 42 items would cost lots of time. Ideally, old datasets from earlier research with the 42 items could be used.

4.5.2. Other Factors affecting Satisfaction

Future research could focus on technology or experience with conversational systems rather than experience with chatbots specifically. In this research, the following two persons would have been grouped into the same category if they both do not use chatbots often: Person A who uses a computer once a month and person B who uses it daily and has lots of experience with technology. While they both do not interact with chatbots often, they would still differ in their level of experience with technology which might affect satisfaction. For example, a person with little technology experience might encounter difficulties when navigating the chatbot. On the other hand, a person with more technology experience can fall back on their general technology knowledge to navigate the chatbot. In previous research, it was shown that users with different levels of internet competency needed different types of

assistance to perform well on a task with conversational agents (Chattaraman et al., 2019). Moreover, students with higher internet competency have also been found to give higher usability ratings to websites (Meiselwitz & Trajkovski, 2006). Thus, instead of focusing on previous experience with chatbots, future research should investigate the relationship between satisfaction and computer or internet competency.

Second, the BUS-11 should be tested with diverse users. Currently, most research is done with university student samples as they are more easily available and large samples are needed to validate the BUS-11. Next to that, the BUS-11 is being tested in different languages (Borsci et al., 2021; Huijsmans, 2022). However, the BUS-11 has not yet been tested with users who have disabilities. This group could be included in future research.

Third, future research should be conducted regarding the relationship between trust and satisfaction. As described previously, research could focus on which features make a chatbot trustworthy and how the trustworthiness of a chatbot affects satisfaction. However, this should be separated from research into the factorial structure of the BUS-11. The reason is that users would have to respond to even more items after each chatbot interaction. This would be tiring and likely lead to dropouts. Therefore, at least the data collection should be split between research into the BUS-11 factorial structure and research into the relationship between trust and satisfaction.

Fourth, future research could control for the task description complexity. For all chatbots, the task description was of similar length. However, some chatbots required a more elaborate explanation and potentially more complex task description. If a task could not be completed, it was mainly because the participant could not find the chatbot. This means it was not about the task itself, but finding the chatbot on the web page. Still, some chatbots needed more attention than others. For example, Chatbot 3 (Dropbox) required less task complexity and information compared to Chatbot 8 (Health Checker) where the participant had to remember and fill in lots of information about their (in this research fictional) symptoms. While this is the nature of the chatbot and it likely had an effect on satisfaction, it could be valuable to control for task complexity in future research. For example, this could be done by adding an item which measures this complexity or by defining the minimum and maximum number of interactions that the participant must perform with a chatbot that is used in the study.

4.8. Conclusion

This research provided new insights into the factorial structure of the BUS-11 from a psychometric and designometric perspective. It validated the previously found five-factor

structure from a psychometric point of view and showed that there likely is a different factorial structure from a designometric perspective. Thus, it provides a first step towards establishing a BUS from a designometric perspective. Moreover, it provides a starting point for designometrics to be more often used in the development of scales. Overall, this research contributed to having a solution for the growing demand for a singular scale that can assess satisfaction with a chatbot – in a world in which chatbots and conversational agents are becoming more and more prevalent in everyday life.

5. References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*(October), 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards Conversational Human-Computer Interaction. *AI Magazine*, *22*(4), 27–38. https://doi.org/https://doi.org/10.1609/aimag.v22i4.1590

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly*, *37*(4), 101490. https://doi.org/10.1016/j.giq.2020.101490

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*(July), 101473. https://doi.org/10.1016/j.tele.2020.101473

Bearden, W. O., & Teel, J. E. (1983). Selected Determinants of Consumer Satisfaction and Complaint Reports. *Journal of Marketing Research*, *20*(1), 21. https://doi.org/10.2307/3151408

Benbasat, I., & Wang, W. (2005). Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, *6*(3), 72–101. https://doi.org/10.17705/1jais.00065

Benke, I., Gnewuch, U., & Maedche, A. (2022). Understanding the impact of control levels over emotion-aware chatbots. *Computers in Human Behavior*, *129*(November 2021), 107122. https://doi.org/10.1016/j.chb.2021.107122

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., Bartolucci, F., 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., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, *26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & van der Velde, F. (2021). *Confirmatory Factorial Analysis of the Chatbot Usability Scale: A Multilanguage Validation*.

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10673 LNCS*(November), 377–392. https://doi.org/10.1007/978-3-319-70284-1_30

Brooke, J. (1996). SUS: A "Quick and Dirty" Usability Scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdemeester (Eds.), *Usability Evaluation In Industry* (pp. 189–194). Taylor & Francis. https://doi.org/10.1201/9781498710411-35

Chang, H. H., & Chen, S. W. (2008). The impact of customer interface quality, satisfaction and switching costs on e-loyalty: Internet experience as a moderator. *Computers in Human Behavior*, *24*(6), 2927–2944. https://doi.org/10.1016/j.chb.2008.04.014

Chattaraman, V., Kwon, W. S., Gilbert, J. E., & Ross, K. (2019). Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior*, *90*, 315–330. https://doi.org/10.1016/j.chb.2018.08.048

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, *117*(November 2017), 587–595. https://doi.org/10.1016/j.jbusres.2018.10.004

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human Computer Studies*, *58*(6), 737–758. https://doi.org/10.1016/S1071-5819(03)00041-7

Cortina, J. M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811–817. https://doi.org/10.1017/S1351324916000243

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11193 LNCS*, 194–208. https://doi.org/10.1007/978-3-030-01437-7_16

Griol, D., Carbó, J., & Molina, J. M. (2013). An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence*, *27*(9), 759–780. https://doi.org/10.1080/08839514.2013.835230

Grunwald, T., & Corsbie-Massay, C. (2006). Guidelines for cognitively efficient multimedia

learning tools: Educational strategies, cognitive load, and interface design. *Academic Medicine*, *81*(3), 213–223. https://doi.org/10.1097/00001888-200603000-00003

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 139-183. https://doi.org/10.1016/s0166-4115(08)62386-9

Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Computers in Human Behavior*, *49*, 245–250. https://doi.org/10.1016/j.chb.2015.02.026

Hoffman, D. L., Novak, T. P., & Peralta, M. (1999). Building Consumer Trust Online. *Communications of the ACM*, *42*(4), 80–85. https://doi.org/10.1145/299157.299175

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huey, B. M., & Wickens, C. D. (1993). Workload transition: implication for individual and team performance. In *National Academy Press*. http://www.dtic.mil/dtic/tr/fulltext/u2/a274538.pdf

Huijsmans, M. (2022). *The Chatbot Usability Scale: An Evaluation of the Dutch Version of the BUS-11* [University of Twente]. http://essay.utwente.nl/90654/

Hussain, S., Ameri Sianaki, O., & Ababneh, N. (2019). A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In *Advances in Intelligent Systems and Computing* (Vol. 927). Springer International Publishing. https://doi.org/10.1007/978-3-030-15035-8_93

International Organization for Standardization. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11 : Guidance on usability*. https://doi.org/10.3403/01879403

Joosten, H., Bloemer, J., & Hillebrand, B. (2016). Is more customer control of services always better? *Journal of Service Management*, *27*(2), 218–246. https://doi.org/10.1108/JOSM-12-2014-0325

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods and Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kokini, C. M., Lee, S., Koubek, R. J., & Moon, S. K. (2012). Considering Context: The Role of Mental Workload and Operator Control in Users' Perceptions of Usability. *International Journal of Human-Computer Interaction*, *28*(9), 543–559.

https://doi.org/10.1080/10447318.2011.622973

Kortum, P., & Johnson, M. (2013). The relationship between levels of user experience with a product and perceived system usability. *Proceedings of the Human Factors and Ergonomics Society*, 197–201. https://doi.org/10.1177/1541931213571044

Larivière, B., Bowen, D., Andreassen, T. W., Kunz, W., Sirianni, N. J., Voss, C., Wünderlich, N. V., & De Keyser, A. (2017). "Service Encounter 2.0": An investigation into the roles of technology, employees and customers. *Journal of Business Research*, *79*, 238–246. https://doi.org/10.1016/j.jbusres.2017.03.008

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE - When there's no time for the SUS. *Conference on Human Factors in Computing Systems - Proceedings*, 2099–2102. https://doi.org/10.1145/2470654.2481287

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLoS ONE*, *13*(8), 1–36. https://doi.org/10.1371/journal.pone.0199661

Longo, L., & Dondio, P. (2016). On the relationship between perception of usability and subjective mental workload of web interfaces. *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, *1*, 345–352. https://doi.org/10.1109/WI-IAT.2015.157

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The Effect of Experience on System Usability Scale Ratings. *Journal of Usability Studies*, *7*(2), 56–67.

Meiselwitz, G., & Trajkovski, G. (2006). Effects of computer competency on usability and learning experience in online learning environments. *Proc. - Seventh ACIS Int. Conf. on Software Eng., Artific. Intelligence, Netw., and Parallel/Distributed Comput., SNPD 2006, Including Second ACIS Int. Worshop on SAWN 2006*, 339–342. https://doi.org/10.1109/SNPD-SAWN.2006.37

Microsoft. (2022). *Microsoft Excel*. https://www.microsoft.com/nl-nl/microsoft-365/excel/

Muir, B. M., & Moray, N. (1996). Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. https://doi.org/10.1080/00140139608964474

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An Initial Model of Trust in Chatbots for Customer Service - Findings from a Questionnaire Study. *Interacting with Computers*, *31*(3), 317–335. https://doi.org/10.1093/iwc/iwz022

Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, *54*(2), 4–8.

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., & Mazurek, G. (2019). In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, *62*(6), 785–797. https://doi.org/10.1016/j.bushor.2019.08.005

Qualtrics. (2022). *Qualtrics XM // The Leading Experience Management Software*. https://www.qualtrics.com/uk/?rid=ip&prevsite=en&newsite=uk&geo=NL&geomatch=uk

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Ren, R., Castro, J. W., Acuña, S. T., & De Lara, J. (2019). Evaluation Techniques for Chatbot Usability: A Systematic Mapping Study. *International Journal of Software Engineering and Knowledge Engineering*, *29*(11–12), 1673–1702. https://doi.org/10.1142/S0218194019400163

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/https://doi.org/10.18637/jss.v048.i02

Schmettow, M. (2021). Psychometrics and design-o-metric models. In *New statistics for Design Researchers: A Bayesian workflow in Tidy R*. Springer. https://doi.org/10.1007/978-3-030-46380-9

Schmettow, M. (2022). *bayr*. https://github.com/schmettow/bayr

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

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Toma, C. L. (2010). Perceptions of trustworthiness online: The role of visual and textual information. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 13–21. https://doi.org/10.1145/1718918.1718923

University of Twente. (2022). *Test Subject Pool BMS*. https://utwente.sona-systems.com/Default.aspx?ReturnUrl=%2F

Wang, W., & Benbasat, I. (2008). Attributions of trust in decision support technologies: A study of recommendation agents for e-commerce. *Journal of Management Information Systems*, *24*(4), 249–273. https://doi.org/10.2753/MIS0742-1222240410

Youn, S., & Jin, S. V. (2021). “In A.I. we trust?” The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging “feeling economy.” *Computers in Human Behavior*, *119*(February), 106721. https://doi.org/10.1016/j.chb.2021.106721

Zamora, J. (2017). *I’m Sorry, Dave, I’m Afraid I Can’t Do That*. 253–260. https://doi.org/10.1145/3125739.3125766

Zijlstra, F. R. H., & Doorn, L. (1985). The construction of a scale to measure subjective effort. *Delft, The Netherlands: Delft University of Technology, Department of Philosophy and Social Sciences, November*.

6. Appendices

Appendix A: BUS-11 and its five factors (5-point Likert scale ranging from strongly disagree to strongly agree)

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 that the chatbot informs me of any possible privacy issues
5. Time response	11. My waiting time for a response from the chatbot was short

Appendix B: Chatbots with the respective URL and task description

Chatbot 1: Adobe
https://helpx.adobe.com/support.html
You are considering a monthly subscription for Adobe Creative Cloud. Your goal is to find out how much the monthly subscription costs for Adobe Creative Cloud. You need to use the chatbot to complete this task.

Chatbot 2: Figma
https://help.figma.com/hc/en-us
Figma is a digital prototyping and brainstorming tool. You have inserted a picture in the online environment, but it is not loading, even after 10min. Your goal is to find out possibilities why the image is not loading. You need to use the chatbot to complete this task.

Chatbot 3: Dropbox
https://www.dropbox.com/support
Dropbox is a platform where you can store files. Your files have now disappeared from the web version of Dropbox, but they are still available on your desktop. Your goal is to find out possibilities why the files are gone. You need to use the chatbot to complete this task.

Chatbot 4: Access Bank
https://www.accessbankplc.com/pages/customer-support.aspx
Access Bank is a commercial bank. You want to buy an expensive item online and need to increase your transfer limit. Your goal is to find out how you can increase your transfer limit. You need to use the chatbot to complete this task.

Chatbot 5: Air New Zealand
https://www.airnewzealand.co.nz/help-and-contact
You would like to know what the flight departures are for all Air New Zealand flights. For that you need to access the flight information tool. You do not have a specific flight number. Your goal is it to be redirected to the website where you can look up all flight departures. You need to use the chatbot to complete this task.

Chatbot 6: Royal Bank of Scotland
https://www.rbs.co.uk/support-centre.html#cora
You have a loan at the Royal Bank of Scotland. As you recently inherited a lot of money, you could pay off the loan all at once now. You are looking for an online form where you can request that. You do not own a smartphone, but a laptop. Your goal is to find the site where you can request to pay your loan all at once. You need to use the chatbot to complete this task.

Chatbot 7: Singapore Airlines
https://www.singaporeair.com/en_UK/sg/support/kris-the-chatbot/
You will fly directly from Amsterdam to Singapore in three days. You will fly with Singapore Airlines in the economy class. You are wondering how much your checked luggage is allowed to weigh. Your goal is to find out what the maximum weight for the checked luggage on your flight is. You need to use the chatbot to complete this task.

Chatbot 8: Buoy Health Checker
https://www.buoyhealth.com/symptom-checker/

Two days ago, your throat started hurting. Now you have a fever of 39.5 °C. You think that you might have the COVID-19 virus. Your goal is to find out if you might have COVID-19. You need to use the chatbot to complete this task.

Chatbot 9: Benefit cosmetics
https://www.benefitcosmetics.com/en-us/contact-us
You have ordered cosmetics to your home address, but they have not shown up yet. It has been 1 month since you placed the order. Your goal is to find out how long it normally takes for Benefit Cosmetics to process an order. You need to use the chatbot to complete this task.

Chatbot 10: Singapore Government
https://www.gov.sg/
You are a long-term resident in Singapore. Your goal is to find out if you have to pay for a COVID-19 vaccine shot in Singapore. You need to use the chatbot to complete this task.