Artificial Intelligence Conversational Agents: A Measure of Satisfaction in Use

Norberts Bekmanis

University of Twente

Faculty of Behavioural, Management and Social sciences

Master Thesis (35 EC) in Human Factors & Engineering Psychology

1st Supervisor:	Dr Simone Borsci	s.borsci@utwente.nl
2nd Supervisor :	Dr Martin Schmettow	m.schmettow@utwente.nl

May 12, 2023

Abstract

As technology evolves and takes over tasks done by humans, conversational agents (chatbots) have been and are implemented into various interactive systems. Federici et al. (2020) have pointed out an increase of necessity in the domain of chatbots for validated scales for human-chatbot interaction measurements. Although existing standardised scales (UMUX, SUS, CSUQ) are usable to measure user satisfaction, they lack the focus on chatbots and their aspects (Tariverdiyeva & Borsci, 2019).

The aim of this study was to advance previous work done on the new scale BUS-11 (Borsci et al., 2022) to investigate the BUS-11 from the psychometrics and designometrics perspectives, and the relationships between BUS-11 and UMUX-Lite, RSME, and familiarity.

In total, there were 98 participants recruited (psychometrics perspective) and 31 chatbots used (designometrics perspective). To analyse the data, a Confirmatory Factor Analysis (CFA) and the extension to a Structural Equation Model (SEM) for regressions between BUS-11 and UMUX-Lite, RSME, and familiarity were performed, as well as Cronbach's alpha was calculated to see the level of reliability.

The results showed that the five-factor structure model of the BUS-11 was confirmed from the psychometrics perspective, however, more research should be done from the designometrics perspective, as the results suggested a different (smaller) factorial structure. The items of the BUS-11 proved to be reliable from both perspectives, and the analysis showed that BUS-11 is a valid measurement tool for user satisfaction of chatbots. Lastly, the results showed a negative correlation between BUS-11 and RSME, and a positive correlation between BUS-11 and familiarity, however, a weak one.

Keywords: chatbots, conversational/virtual agents, user satisfaction/experience, interaction, Bot Usability Scale (BUS-11)

Table of Contents

Introduction

As technology evolves and takes over tasks done by humans, conversational agents have been and are implemented into various interactive systems, both online and offline, such as websites, phone apps, social networks, and media, as well as cars etc. Those agents work as customer support/service to aid users' needs, and interactions with these systems and agents create and affect user experience (UX). Conversational agents are text/voice interfaces, virtual agents, robots, or as most used and called – chatbots (Borsci et al., 2021, Rajaobelina, Prom Tep, Arcand, & Ricard, 2021). Conversational agents and chatbots sometimes have different definitions, however, often in literature both are referred to as synonymous (Borsci et al., 2021; Io & Lee, 2017; Vaidyam, Wisniewski, Halamka, Kashavan, & Torous, 2019). In the present study, we will use the term chatbot to refer to all those virtual agents that can communicate with users in the text and/or voice by simulating and using natural language based on Artificial Intelligence (AI) to imitate human interactions as best as possible (Borsci et al., 2021).

The general aim of this study is to advance previous work done on a new scale (BUS-11, see Appendix D) to assess satisfaction with chatbots, with a focus on the assessment of chatbots to further streamline the reliability and validity of the scale. In addition, to investigate the relationships between such satisfaction scale and UMUX-Lite, RSME (mental effort), and familiarity (previous experience).

Previously, such rating scales have been put into psychometrics perspective, however, measuring a person with a rating scale is different from measuring a design. This is what Schmettow and Borsci (2020) calls the *psychometric fallacy* – validating designometrics rating scales as if they were psychometric. Therefore, these two perspectives – psychometrics and designometrics – will be explored and compared in this study.

A brief history of chatbots

In 1950, Alan Turing devised a test in which a computer program (a machine) had a text-only conversation with a human evaluator. During the test, the evaluator has natural language conversations with another human and a machine, all three parties separated from each other. The evaluator would have to judge these conversations and be able to distinguish between the human and the machine. Although the Turing test was designed to test the intelligence of the machine and not the conversation, it is considered by many to be the generative idea of chatbots (Adamopoulou & Moussiades, 2020; Turing, 1950).

Years later, in 1966, Joseph Weizenbaum created the first chatbot - ELIZA. This chatbot represents the role of a Rogerian psychotherapist by asking and answering open-ended questions (Adamopoulou & Moussiades, 2020; Klopfenstein, Delpriori, Malatini, & Bogliolo, 2017; Zemčik, 2019). In short, Rogerian psychotherapy (also known as client/person-centred therapy) consists of three main elements: empathy, congruence (or authenticity), and unconditional positive regard (Hopper, 2021). ELIZA's knowledge is limited and domain-specific, so it cannot sustain long discussions and learn from them. However, despite its limitations, ELIZA was a source of inspiration for the development of chatbots (Adamopoulou & Moussiades, 2020; Klopfenstein et al., 2017; Zemčik, 2019).

The next chatbot – PARRY by Kenneth Mark Colby – came to life in 1972. PARRY was designed to play the opposite role of ELIZA - a patient with schizophrenia. Compared to ELIZA, PARRY was designed to have an advanced control structure and a more defined personality. Despite the enhancements, PARRY was still slow to respond and had limited ability to understand and express concepts (Adamopoulou & Moussiades, 2020).

In 1991, another chatbot is mentioned – Chatterbot. It was an artificial player created in an online MultiUser Dungeon (MUD) called TINY-MUD, where players could create and personalise their areas in a virtual world. This world was filled with players who communicate by typing, providing a nice opportunity for developers to test conversational bots. Chatterbot could not only communicate but also explore and discover the world, as well as participate in a multiplayer card game “Hearts”, if wanted/requested. However, of course, the primary goal was to answer questions and maintain appropriate responses. Chatterbot was successful because players assumed that everyone was a real person unless Chatterbot made a significant mistake that raised suspicion. Interestingly, many of the players preferred to interact with Chatterbot rather than with other players (Adamopoulou & Moussiades, 2020; Mauldin, 1994).

A year later, in 1992, another advancement in chatbot development was created - speech synthesis using Creative Labs' Sound Blaster sound cards. The name of the program itself was Dr. Sbaitso (Sound Blaster Artificial Intelligent Text to Speech Operator). This allowed the operator (chatbot) to communicate verbally, making the chatbot more human. However, it was only able to converse in a simple way (Adamopoulou & Moussiades, 2020; Zemčik, 2019).

The next step forward in chatbots was the first online chatbot called ALICE (Artificial Linguistic Internet Computer Entity), which was inspired by ELIZA. It used a new language - Artificial Intelligence Markup Language (AIML) - which is the main difference between ALICE and ELIZA. The ELIZA chatbot had 200 keywords and rules, while ALICE had 41,000 templates and related patterns. ALICE was a significant improvement, but it did not have intelligent features and could not express emotions or attitudes (Adamopoulou & Moussiades, 2020).

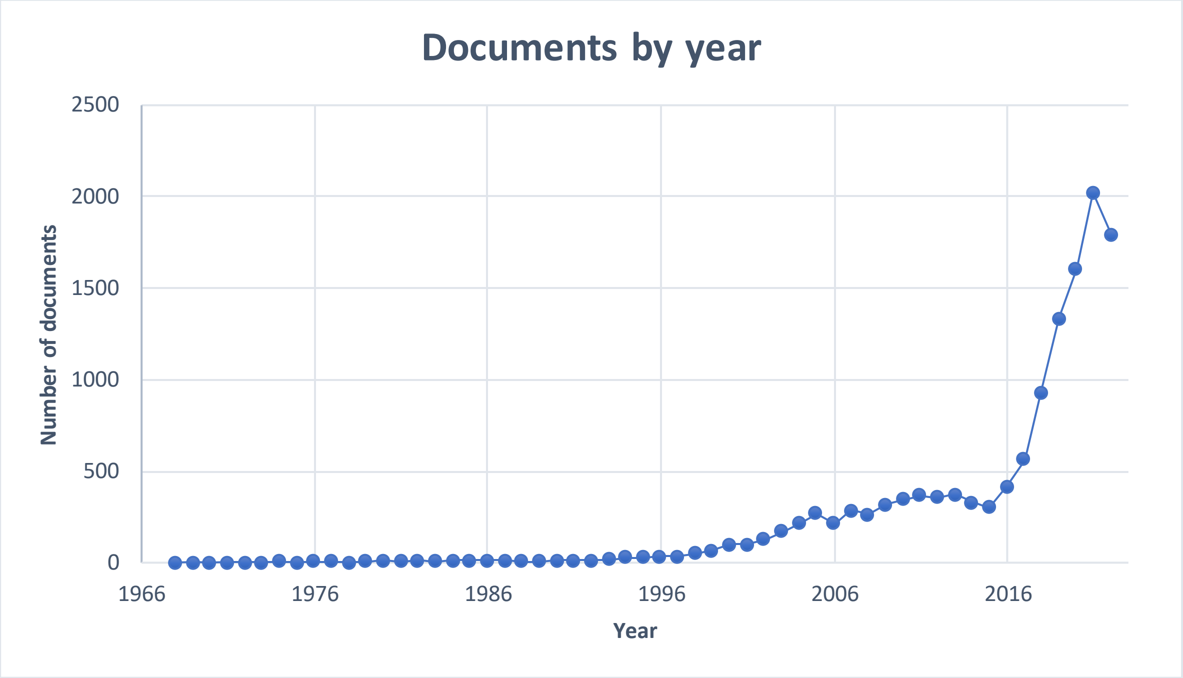
The development of chatbots over these years has been significant, but one of the big leaps was the creation of a chatbot called SmarterChild in 2001. This was the first chatbot that could help humans with tasks, thanks to its ability to retrieve information from databases – such as news, weather, sports, etc. It was integrated into messengers such as Microsoft (MSN) and America Online (AOL). SmarterChild boosted the development of AI and Human-Computer Interaction (HCI) as information retrieval could be done by communicating with a chatbot (Adamopoulou & Moussiades, 2020; Molnár & Zoltán, 2018).

The next step into the development came with the creation of voice assistants. They are built into various devices (e.g., smartphones) that can understand voice commands, communicate through digital voices, and handle some tasks, for example, at home by monitoring other automated devices. The most popular voice assistants are Amazon Alexa, Apple Siri (Siri), Google Assistant, IBM Watson, and Microsoft Cortana (Adamopoulou & Moussiades, 2020; Hoy, 2018).

The somewhat final leap in the evolution of AI and chatbots took place in 2016 when the communication between people and manufacturers changed. Meaning, chatbots were created for brands and/or services on social media platforms, allowing users to perform everyday actions within their messaging applications. By the end of the year, around 34,000 chatbots were being used to perform a variety of tasks and actions within areas such as education, healthcare, marketing and others (Adamopoulou & Moussiades, 2020; Dale, 2016; Powton, 2018). Figure 1 shows how many documents related to chatbots can be found in Scopus by year, and supports that 2016 was the beginning of the rise of chatbot development, use, and the peak of interest from companies and stakeholders.

Figure 1

Number of documents by year in search results on Scopus from the year 1968 to 2022 (20.12.2022)



Note. Including keywords "chatbot", "conversation agent" "conversational agent" or "conversational interface"

In total, there are 13,286 documents. During the years 2006-2016, there were, approximately, 250-350 documents per year, and the number of documents has increased significantly since 2016 (416 documents) and peaking the highest in 2021 (2023 documents) (Scopus, n.d.).

The global objective for the creation and development of chatbots has been and continues to be to improve the quality of customer service and experience (Rajaobelina et al., 2021; Thomaz, Salge, Karahanna, & Hulland, 2019). A systematic literature review by Bavaresco et al. (2020) examined where and how chatbots are being used, and showed that approximately 40% of the studies in this review are in the commerce domain and that chatbots are mostly used for customer support. To continue, they are more appealing to users' needs when compared to, for example, FAQs, which are static and unengaging. Meanwhile, chatbots can provide comfort and efficiency in their interactions with users (Følstad, & Brandtzæg, 2017; Ranoliya, Raghuwanshi, & Singh, 2017). In addition, users often perceive a chatbot not only as an assistant, but more as a friendly companion, as according to the findings of Xu, Liu, Guo, Sinha, and Akkiraju (2017), 40% of user queries are more emotional than informative.

The evolution of chatbots has come a long way, and today the interaction between humans and chatbots is very different from the capabilities of their predecessors. Although the main difference between a human and a chatbot is the perception of empathy, chatbots are becoming more aware of users' feelings. Now, chatbots can act more like humans, even sharing personal thoughts or describing some dramatic events, making it harder to distinguish between humans and chatbots (Adamopoulou & Moussiades, 2020; Fernanded, 2018; Shah, Warwick, Vallverdú, & Wu, 2016).

Differences and categories of chatbots

To better explain the variety of systems that fall under the term *chatbot,* chatbots can be distinguished depending on their design, goals, ways of interacting with users, as well as categories. One of the differences, regarding the design, lies in whether chatbots are driven by machine learning or by algorithms. This affects their communication with the users, which can be textual, audial, or sometimes interacting with images. Another difference between the chatbots is their core function, which could be one of the main distinctions between them - what are their capabilities, what can they assist with, as their goals may also vary. Chatbots can simply provide information to the user or have simple interactions with the user. However, they can also assist the user to perform tasks on their own or together by asking the user to perform certain steps if the chatbot is unable or incapable of performing certain tasks (Adamopoulou & Moussiades, 2020). Today, people encounter and interact with different types of chatbots available on the market, and the following table provides an overview of chatbot categories and types.

Table 1

Chatbot categories and types (Adamopoulou & Moussiades, 2020)

Categories	Types
Knowledge domain	Generic
	Open domain
	Closed domain
Service provided	Interpersonal
	Intrapersonal
	Inter-agent
Goals	Informative
	Chat based/Conversational
	Task based
Response Generation Method	Rule based
	Retrieval based
	Generative
Human-aid	Human-mediated
	Autonomous
Permissions	Open-source
	Commercial
Communication channel	Text
	Voice
	Image

As established, chatbots vary in their purposes, functions, capabilities, and categories, and they can be very helpful to users. The user perspective side is that chatbots enhance the quality of customer service and experience due to their accessibility, convenience, and efficiency (Ling, Tussyadiah, Tuomi, Stienmetz, & Ioannou, 2021; Rajaobelina et al., 2021; Thomaz et al., 2019). However, the user perspective is just one of the sides. To understand the importance of chatbots, it is also important to understand the contribution of chatbots to the companies that provide them and the companies and services that implement them for users. The better chatbots can support the decision-making and information retrieval of users, the better the user experience. For businesses, this means increased user satisfaction, association and relationship with the business, its services and/or products. In addition, chatbots have a positive impact on operational costs and can provide 24/7 assistance to users for a variety of activities/tasks (Borsci et al., 2021; Følstad, Skjuve, & Brandtzaeg, 2019; Paikari & van der Hoek, 2018). This reduces the workload on humans and allows them to focus on other priorities for the business, while allowing chatbots to take over these processes and tasks.

Businesswire (2019) forecasted that by the year 2020 chatbots will handle 85% of user interactions. Furthermore, Global Market Insights has suggested that the overall market size for chatbots will exceed USD 1.3 billion globally by 2024 (Park, Ahn, Thayisay, & Ren, 2019; Rajaobelina et al., 2021). However, the market was worth almost USD 1 billion in 2017, and Businesswire (2019) suggested that it will be worth much more – just over USD 5.5 billion by 2023.

Evaluation and measures of chatbots

After researching what is a chatbot, its origins, and the progress of development, it is also worth exploring the evaluation of chatbots, as well as the interaction between a user and a chatbot. Federici et al. (2020) have pointed out an increase of necessity in the domain of chatbots for validated scales for human-chatbot interaction measurements, so that developers can improve their chatbots already during the design phase. To better understand the evaluation and the interaction, it is beneficial to explore the qualities of a chatbot that affect the ease of use of the chatbot and its usefulness for the user in order to achieve their intended goals (Bevan, 1995). For that, first, a term ‘usability’ must be defined, and the definition by ISO 9241-11:2018 – “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO, 2018). As visible, the qualities of a chatbot are measured by usability.

Overall, there are multiple measurements (scales) to assess perceived usability in a reliable manner, ranging from short and quick to long scales. The shortest ones are the four-items Usability Metric for User Experience scale (UMUX; Finstad, 2010) and its shortened two-items version (UMUX-Lite; Lewis, Utesch, & Maher, 2013). Their difference is that UMUX-Lite uses only two positive statements, whereas UMUX has both positive and negative statements. UMUX-Lite covers whether the system’s capabilities have met the requirements of the user and the ease of use of the system. Various research regarding the psychometric properties of this scale have shown an acceptable reliability with estimates of Cronbach's α between 0.77 and 0.86, as well as an acceptable validity and sensitivity (Berkman & Karahoca, 2016; Borsci, Federici, Bacci, Gnaldi, & Bartolucci, 2015; Lewis et al., 2013; Lewis Utesch, & Maher, 2015).

Other scales that are considered to be one of the most popular ones to measure perceived usability are System Usability Scale (SUS) and Computer System Usability Questionnaire (CSUQ). SUS consists of 10 statements (five positive and five negative) that are measured on a 5-point Liker scale (Brooke, 1996). CSUQ is a little longer, consisting of 16 statements, however, split into three subcategories – System Usefulness, Information Quality, and Interface Quality) –, and measured on a 7-point Liker scale (Lewis, 2002). The reliability of these two scales with Cronbach's α is even better – 0.97 and 0.93 respectively (Lewis, 2018).

However, as previously stated, usability includes "satisfaction in a specified context of use", therefore, the usability of a chatbot is also linked to another aspect – user satisfaction. ISO 9241-11:2018 defines user satisfaction as the "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (ISO, 2018). Even though the previously mentioned standardised scales are also usable to measure user satisfaction, they lack the focus on chatbots and their aspects (Tariverdiyeva & Borsci, 2019). In addition, these scales can only show the general satisfaction of a user without a deeper insight into specific aspects of a system (chatbot) (Balaji & Borsci, 2019; Tariverdiyeva & Borsci, 2019). Considering this, as well as the mutual agreement with other experts and end-users about the importance of user satisfaction, Balaji and Borsci (2019) created the user satisfaction questionnaire (USQ), consisting of 42 items measuring user satisfaction after an interaction with a chatbot. They based the questionnaire on Tariverdiyeva and Borsci (2019) identified 27 features for the perceived usability of chatbots, as well as on consequential literature review and on feedback from a focus group.

Bot Usability Scale (BUS)

Nevertheless, there were no standardised tools that assess user satisfaction and the end-user's perception of the quality of interaction with chatbots, as well as no previous studies that have identified and tested a model for that to develop a reliable tool for chatbot designers (Borsci et al., 2021). That was until Borsci et al. (2021) aimed to provide a toolkit to aid the chatbot designers during the development of chatbots so that the needs of the end-users, as well as the quality of interaction with chatbots, can be considered and assessed beforehand. To achieve this, four sequential studies were performed:

1. An examination of chatbot attributes identified by Radziwill and Benton (2017), and a systematic review that was also based on USQ (42 items) by Balaji and Borsci (2019).
2. An online survey with chatbot designers and end-users to reach an agreement on the list of attributes.
3. An expansion of the list attributes, a creation of a list of items for a questionnaire, and focus group sessions to develop the first version of the new scale called The Bot Usability Scale (BUS).
4. A pilot and an analysis of the scale and its psychometric properties to develop it into a final version.

So, the BUS scale was developed and tested using an exploratory factorial analysis of the USQ model that, essentially, was reduced to 15 items with five factors (BUS-15) with an overall reliability of 0.87. It and its factors also strongly correlate with UMUX-Lite (between 0.61 and 0.87), indicating reliability in assessing the end-user's overall satisfaction, as well as adding new aspects that are not considered by classic non-chatbot satisfaction scales. Borsci et al. (2021) recommended to test and validate BUS-15 further, and it was by Borsci, Schmettow, Malizia, Chamberlain, and van der Velde (2022b) and their multilanguage validation of BUS-15 together with UMUX-Lite. They were also translated in Dutch, German, and Spanish, and after the gathered data, a shorter and even more reliable and better version of BUS was created, consisting of 11 items (BUS-11). The results from these different languages have shown that they were not significantly different from the original version (English), therefore, indicating that this scale can be translated in used in various languages. In addition, it also has a positive correlation with UMUX-Lite scale. This improvement from BUS-15 to BUS-11 could be an even better way for the assessment of chatbots. With Borsci et al. (2022b) words – "this tool could be a way to harmonise and enable comparability in the field of human and conversational agent interaction." (p. 1). For this study, BUS-11 (see Appendix D) will be used.

However, before heading to the aim of this study, another term and section are worth exploring. Previously, we have established the importance of perceived usability and user satisfaction when interacting with chatbots. Interaction with chatbots creates user experience (UX), and perceived usability and user satisfaction are part of it. The definition of UX by ISO (2018) – "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service". The previously mentioned scale (BUS) is a measure of perceived usability (satisfaction), which is a fundamental component/metrics of the assessment of usability and user experience. To continue, we need to see what makes UX and how it is influenced, as during the creation of BUS factors affecting the UX in the following section were explored and incorporated (Borsci et al., 2021).

Factors affecting UX

Although the chatbot industry has grown significantly and conversational solutions are becoming popular, there are issues with chatbots that must be considered. Aside from the chatbots themselves and the ways of their evaluation, it is also important to research what should be considered during the development of chatbots from the UX perspective. Based on the literature, five factors have been researched – productivity, humanisation of chatbots, creepiness, trust, and familiarity. However, only familiarity will be used in this study for the reasons mentioned below.

Familiarity

Gefen's (2000) translation from German of Luhmann's (1979) definition of familiarity is as follows – "an understanding, often based on previous interactions, experiences, and learning of what, why, where and when others do what they do" (p. 3/727). Perhaps it is a broad definition covering aspects of someone's behaviour, but it can be well linked to technology and chatbots. People interacting with chatbots increase their experience and, hence, familiarity with chatbots and technology in general, and, from the other side, the higher the level of skills with technology (e.g., internet skills), the more useful and less time-consuming users perceive chatbots (Ben Mimoun, Poncin, & Garnier, 2017). Familiarity with technology and chatbots helps in understanding how to use one, so, these prior experiences will also make the future interactions better, as the user is more knowledgeable, more aware of chatbot's capabilities, which also increase the likeliness of higher user satisfaction. These experiences are considered to be more reliable than indirectly gained information, as the user has more information about the functionality (Ben Mimoun et al., 2017; McKnight, Cummings, & Chervany, 1998). Although experiences with chatbots increase familiarity, experience can also be gained throughout knowledge and expertise of chatbots (Ben Mimoun et al., 2017), for instance, reading about chatbots and their capabilities.

Familiarity can be connected to the previously mentioned factors, and they can be to other factors. Productivity – chatbot's efficiency in tasks both in performance and quickness in comparison to the user (Jenkins et al., 2007; Zamora, 2017). By the description of familiarity, the user with a higher level of familiarity towards chatbots will also be able to the chatbot's productivity more efficiently. To continue, people perceive chatbots also as creepy (Ashfaq, Yun, Yu, & Loureiro, 2020), and that perception is influenced by privacy concerns (trust) (Inman & Nikolova, 2017; Ischen, Araujo, Voorveld, van Noort, & Smit, 2020; Mani & Chouk, 2017). When users get more familiar with chatbots, trust towards chatbots also increases, so, familiarity is a precondition for trust in chatbots (Gefen, 2000). Studies by Kortum and Johnson (2013), Lindgaard and Dudek (2003), McLellan et al. (2012), and Sauro and Lewis (2011) have also shown a correlation between user satisfaction and user's experience with a product. Users with a higher level of experience of a product assessed their level of satisfaction as better and higher than the users with a lower level of experience, though System Usability Scale (SUS) was used in these studies. In addition, familiarity also affects the assessment of chatbots (Ben Mimoun et al., 2017).

Hence, considering all this, only familiarity will be used in this study, regarding factors that affect UX. Although Pollmann and Borsci (2021) has also explored the effect of familiarity on satisfaction ratings and stated that it is a non-significant predictor, as well as Huijsmans and Borsci (2022) stated that there is no effect of familiarity on user satisfaction, however, Braun and Borsci (2023) found a positive relationship, it can be tested again, as it is also only a minor part of this study.

Mental workload

Associated to the satisfaction of the users, another important aspect that might affect the interaction with chatbots is mental workload. Meaning, its potential effect on user satisfaction. Mental workload (also known as cognitive workload) can be defined as the amount of resources required for tasks, and the amount required to perform them (Hoedemaeker, 2002). Longo (2018) has proposed a model, which describes the link among mental workload, usability, and objective performance. By this and referring to the definitions of usability and user satisfaction by ISO (2018), in which usability and user satisfaction are linked, it can be proposed that the higher level of mental workload, the lower is user satisfaction measured by standardized scales, therefore, also BUS-11. This has also been proven in the field of e-commerce by Schmutz, Heinz, Métrailler, and Opwis (2009), however, this negative correlation should also be tested with chatbots, as it also one of the sub-goals of this study.

To test that, a measure of mental workload that is reliable and valid is needed. The study by Alimohammadi, Zabiholah, Rahmani, Parsazadeh, and Yeganeh (2019) aimed to determine the validity and reliability of three different measures of mental workload to – the Rating Scale Mental Effort (RSME), Integrated Workload Scale (IWS), and Overall

Workload Scale (OW) in Iran. This study was conducted on 100 male students at Iran University of Medical Sciences. These scales compared similarly regarding the internal validity; however, they presented the RSME as the most reliable one. In addition, this scale is a one-item survey, which would work well in this study, as participants will be asked to rate their mental effort multiple times. Thus, this measure of mental workload is chosen for this study to test the correlation between mental workload and the user satisfaction in the context of chatbots.

Aim of the study

The study aims to re-test the five-factor structure of the BUS-11 (see Appendix D) using other chatbots (in Dutch and/or English), and to assess the reliability of the scale, its internal validity by performing confirmatory factor analysis, and its external validity by using the UMUX-Lite scale. The acquired data in this study will be explored from both psychometrics and designometrics perspectives.

Psychometrics and designometrics perspectives

Multi-item validated scales, generally, serve for one of two reasons – to find out more about the users/participants or about the quality of designs (of systems). In design research, when it comes to valid and reliable measurements of a person, their characteristics, performance etc., a psychometrics model is used. For the quality of designs (aesthetics of a system, system’s capabilities to satisfy the users etc.), another model is necessary – a designometrics model. Those models have some overlaps, but their brief definitions and difference are as follow:

1. The psychometrics perspective measures personal attributes when dealing with persons and items.
2. The designometrics perspective measures design attributes when dealing with designs, persons, and items (Schmettow, 2020; Schmettow & Borsci, 2020).

To continue, in psychometrics studies persons are compared, but in designometrics studies the focus is on designs, so their goal of measurements differs. By their definitions, the psychometric perspective is a two-dimensional matrix (persons by items), in which a large sample of persons is needed, whereas the designometrics perspective is a three-dimension matrix (persons by items by designs), in which a large sample of designs is needed. The basis of the psychometrics matrix is used to create the designometrics matrix. Meaning, the designometrics perspective includes the psychometrics perspective's two-dimensional matrix (encounter), and, therefore, it is possible to "create a psychometric response matrix from a designometric response cuboid, by averaging over designs. At the same time you can create a designometric response matrix from the designometric response cuboid by averaging over persons." (Schmettow, 2020; Schmettow & Borsci, 2020).

The acquired data from this study will show the different attributes of people, such as age, nationality, their level of English, and previous experience (familiarity) with chatbots, hence, the psychometrics perspective comes in handy to see the influence of these attributes on the user satisfaction with chatbots. However, as that is one of the goals of this study, it is more important to compare chatbots and to see whether the BUS-11 is a reliable and valid scale for the assessment of chatbots from another perspective. For that, it is necessary for participants to assess items for multiple designs, therefore, the designometrics perspective is needed.

Research questions

This study aims to investigate the BUS-11 from the psychometrics and designometrics perspectives, and the relationships between such satisfaction scale and UMUX-Lite, RSME (mental effort), and familiarity (previous experience). In line with that, based on previous literature, four research questions are formulated:

1. Can the five-factor structure of the BUS-11 be confirmed from the psychometrics and designometrics perspectives?
2. Are the items of the BUS-11 reliable?
3. Do the results of the BUS-11 correlate positively with the results of the UMUX-Lite?
4. How does mental effort and familiarity (previous experience) of chatbots affect the user satisfaction?

Other studies (see Huijsmans & Borsci, 2022; Braun & Borsci, 2023) attempted similar research with different chatbots (in Dutch and English, respectively). In the present report, the data of such previous studies from the designometrics perspective will be combined with the new generated data in our investigation. In addition, it is worth mentioning, after the analysis of their data, it was noted that there might have been confusion from participants' side with the item one from the BUS-11, especially, in the Dutch version of BUS-11. A similar result was also discovered in the Italian version of the scale (Borsci, Prati, Federici, Malizia, Schmettow, & Chamberlain, 2022). This resulted in the suggestion to slightly change the description of that item. In this study, for the first time in English, the adapted description of that item is included. The difference in descriptions of Item 1 can be seen in the Appendix D.

Methods

Participants

A total number of 98 participants have been recruited by convenience and snowball sampling (participants help recruit other participants), by University of Twente Sona Systems Psychology Test subject pool, in which students of the university can sign-up for studies, and researchers use their own experience and knowledge for participant selection (Everitt & Skrondal, 2010). The inclusion criteria for participants were to be at least 18 years old and to understand English to be able to successfully do the experiment and fill out the questionnaires.

Materials

In this study, an online test to interact and assess chatbots was developed using an online software package Qualtrics. The test was divided into two parts. The first part consists of an introduction to the goal of the study, the Informed Consent (see Appendix A), questions about their demographic characteristics, level of English, previous experience with chatbots, as well as instructions regarding the test (see Appendix B). In the second part, participants were asked to interact and achieve tasks with six chatbots randomly extracted from a database of 9 chatbots (see Appendix C). After the interaction with each chatbot, the participants were asked to assess their experience by answering 15 questions in total: 1 item regarding the completion of the tasks, 11 items about the BUS (see Appendix D), 2 items from Usability Metric for User Experience Lite (UMUX-Lite) (see Appendix E), and 1 item of Rating Scale Mental Effort (RSME) (see Appendix F). Items by BUS and UMUX-Lite were measured on a 5-point Likert scale from Strongly Disagree to Strongly Agree. The experimental part – scenarios, tasks, and links (see Appendix C) – were written and created in the same questionnaire. Besides these online materials, participants need a device (phone, tablet, computer) that is connected to the Internet. For the data analysis, the programming software R (v4.2.3; R Core Team, 2022) was used (see Appendix G for the code used in R).

Procedure

As the experiment is created and carried out using an online software package Qualtrics, everything takes place online. After a participant receives the link to the study, they are briefly informed about the experiment and are asked to read and sign the consent form. Next, they answer questions regarding their demographic characteristics, level of English, and previous experience with chatbots, and read an explanation on how to proceed with the experiment. In the experiment part, the participant is given a scenario and tasks to complete on a website, on which interaction with a chatbot takes place. When the tasks are completed (or not, if for some reasons were not possible to complete), the participant goes back to the questionnaire to answer questions about that chatbot on the BUS scale, user experience (UMUX-Lite) scale, and their effort put to complete the tasks on the RSME scale. The participant does this experiment's part procedure six times in total with six different websites/chatbots that are randomly extracted from a database of 9 chatbots. Afterwards, that is the end of the experiment and they successfully participated in the study.

The experiment and its procedure were approved by the Ethical Committee of the University of Twente (Request nr. 211448).

Data analysis

The obtained data was exported from Qualtrics and transferred into Microsoft Excel as comma-separated values. Then the dataset was cleaned from entries if a participant, for example, could not find a chatbot (by indicating that in the survey) and from incomplete surveys in general. However, if a participant did not fully complete the survey but completed at least one chatbot and corresponding questions, their data were included (86 completes, 12 incompletes but usable), resulting in 520 usable observations in total. The scores on the 5-point Likert scales (Previous experience, BUS-11, UMUX-Lite) and the scores on the 6-point scales (Level of English, Frequency of chatbot usage) were converted to a 0-100 points scale bases on the method by Lewis & Sauro (2020), and the scores on the RSME were converted from a 150-point scale to a 0-100 points scale as well. The dataset was rearranged to be compatible with R (v4.2.3; R Core Team, 2022) so the data was imported into R. In addition, the data for the designometrics perspective were combined with Huijsmans and Borsci (2022) and Braun and Borsci (2023) to increase the sample size of chatbots from 9 to 31 chatbots.

Hence, the psychometrics model is structured as 520 observations with the scores of the previously mentioned scales, and the designometrics model is structured as 31 chatbots with the scores of the BUS-11 scale, averaging by items.

To perform a confirmatory factor analysis (CFA), the CFA function from the R package ‘Lavaan’ (Rosseel, 2021) was used. To assess the BUS-11 factor structure, determine the acceptance of the model, and to further validate the factorial structure, the criteria by Hu and Bentler (1999) were used, which were also used by Borsci et al. (2021):

Table 2

The criteria by Hu and Bentler (1999)

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

To measure the internal consistency of BUS-11, Cronbach’s alpha was calculated to see the level of reliability. The criteria for good reliability was set to ≥ 0.7 (Cortina, 1993; Taber, 2017).

To test the validity of the scale (BUS-11) and the correlations of it with mental effort (RSME) and familiarity (previous experience), the CFA model was extended to a Structural Equation Model (SEM) by adding regressions. Hence, regressions between BUS-11 and UMUX-Lite, BUS-11 and RSME, and BUS-11 and familiarity were performed.

Results

Descriptive statistics

Of those 98 participants (Age – M_{age} = 22.3, SD_{age} = 5.5, Min_{age} = 18, Max_{age} = 52), 38 are male and 60 are female. Their nationalities are Dutch (30), German (35), Latvian (13), and other nationality (20). Their average level of English (measured in levels from A1 to C2) is high (M_{eng} = 74.7, SD_{eng} = 18.5), their average familiarity (previous experience) with chatbots is also high (M_{fam} = 67.8, SD_{fam} = 17.9), and the frequency of chatbot usage is low (M_{freq} = 19.2, SD_{freq} = 11.1), in which 15 participants indicate that they never use chatbots, 74 use rarely, 7 use once per week and 2 use 2-3 times per week.

Abbreviations

The following tables show the names of the BUS-11 items, factors, and their abbreviations that are used in this report. For more information, see Appendix D.

Table 3

Five factors of the BUS-11 and their abbreviations

Abbreviation	Factor
ACF	1 – Perceived accessibility to chatbot functions
QCF	2 – Perceived quality of chatbot functions
QCI	3 – Perceived quality of conversation and information provided
PS	4 – Perceived privacy and security
TR	5 – Time response

Table 4

11 items of the BUS-11 and their abbreviations

Abbreviation	Item	Belonging factor
DET	1 – The chatbot function was easily detectable (e.g., the possibility to modify the settings of the chatbot, make the avatar visible or not etc.).	ACF (1)
FND	2 – It was easy to find the chatbot.	ACF (1)
COM	3 – Communicating with the chatbot was clear.	QCF (2)
TRCK	4 – The chatbot was able to keep track of context.	QCF (2)
DIG	5 – The chatbot's responses were easy to understand.	QCF (2)
ASST	6 – I find that the chatbot understands what I want and helps me achieve my goal.	QCI (3)
aINFO	7 – The chatbot gives me the appropriate amount of information.	QCI (3)
nINFO	8 – The chatbot only gives me the information I need.	QCI (3)
ACC	9 – I feel like the chatbot's responses were accurate.	QCI (3)
PVCY	10 – I believe the chatbot informs me of any possible privacy issues.	PS (4)
TIME	11 – My waiting time for a response from the chatbot was short.	TR (5)

Confirmatory factor analysis

Psychometrics perspective

To test the five-factors structure of the BUS-11 from the psychometrics perspective, a confirmatory factor analysis was performed. The following results are assessed using the criteria by Hu and Bentler (1999) described in Data Analysis to determine the acceptance of the model:

Table 5

The results of the CFA for the criteria by Hu and Bentler (1999)

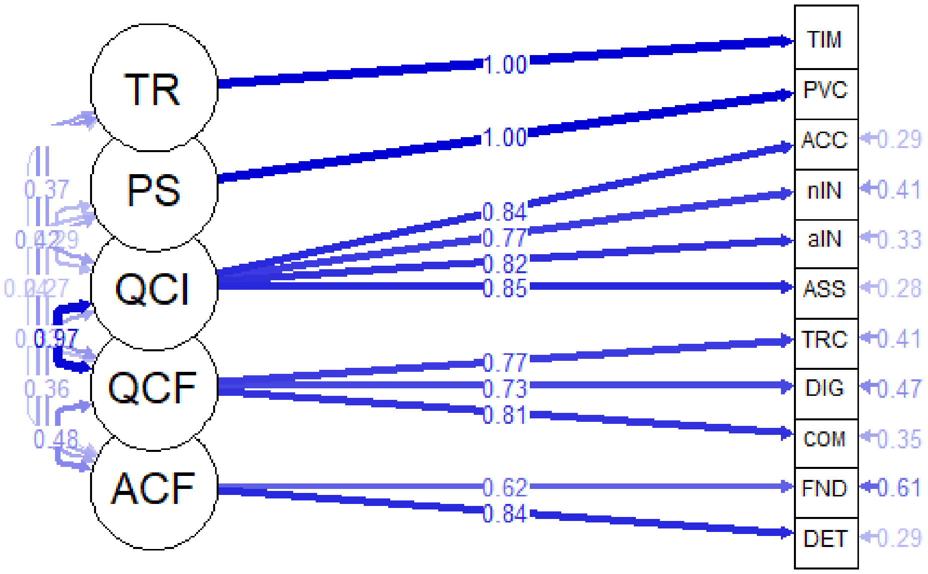
Comparative Fit Index (CFI)	0.980
Tucker-Lewis Index (TLI)	0.970
Root Mean Square Error of Approximation (RMSEA)	0.054
Standardized Root Mean Square Residual (SRMR)	0.027

Both CFI and TLI are greater than 0.95, indicating a good fit. The RMSEA is lower than 0.08, also indicating a good fit, and the SRMR is lower than 0.06, indicating a good fit as well. Thus, the model from the psychometrics perspective is acceptable.

Regarding the factor loadings of the BUS-11, a visualisation of the BUS-11 factors structure from the psychometrics perspective can be seen in the Figure 2. They vary from 0.62 (ACF-FND) to 1. Factors that consist of multiple items (ACF, QCF, QCI) have their loadings from 0.62 to 0.85, while single item factors (PS, TR) loadings are 1. Besides the lowest loading of the FND (0.62), all items explain at least 73% of the variance in each factor, and the total mean value of the variance in each factor explained is 82.3%.

Figure 2

Visualisation of the BUS-11 factors structure (psychometrics perspective)



The correlations between the factors can be seen in Table 6. The correlation range is from -0.001 to 0.823, with the weakest correlation, which is negative, however, almost zero, being between factors 4 (PS) and 5 (TR), and the strongest correlation between factors 2 (QCF) and 3 (QCI). Except these two correlations, all correlations are rather weak, and the mean value between all factors is 0.318.

Table 6

Correlation of the BUS-11 factors (psychometrics perspective)

	ACF	QCF	QCI	PS	TR
ACF					
QCF	0.355**				
QCI	0.266**	0.823**			
PS	0.268**	0.248**	0.274**		
TR	0.228**	0.378**	0.344**	-0.001	

Note. *p<0.05, **p<0.01

Designometrics perspective

To test the five-factors structure of the BUS-11 from the designometrics perspective, a confirmatory factor analysis was performed. The following results are assessed using the criteria by Hu and Bentler (1999) described in Data Analysis to determine the acceptance of the model:

Table 7

The results of the CFA for the criteria by Hu and Bentler (1999)

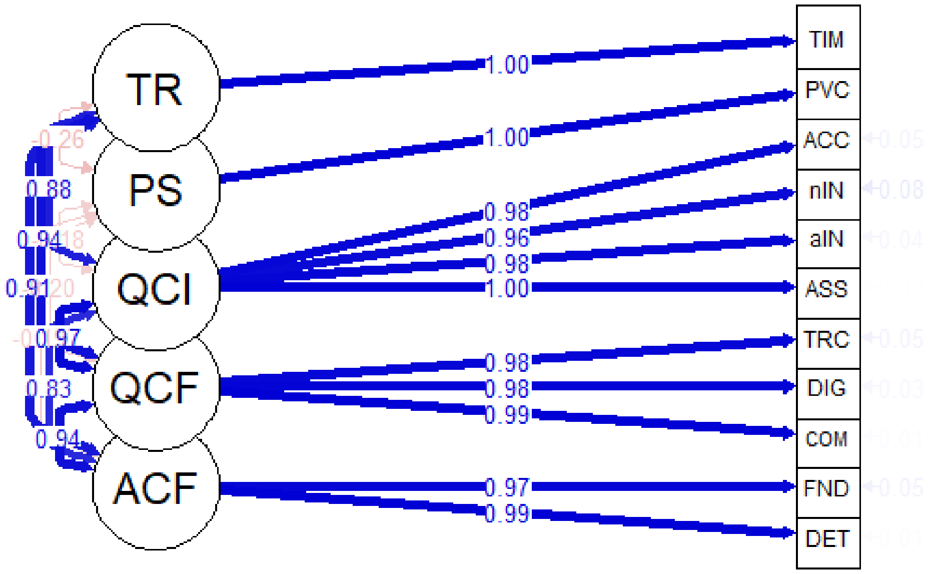
Comparative Fit Index (CFI)	0.904
Tucker-Lewis Index (TLI)	0.853
Root Mean Square Error of Approximation (RMSEA)	0.270
Standardized Root Mean Square Residual (SRMR)	0.028

Both CFI and TLI are lower than 0.95, indicating an insufficient fit. The RMSEA is greater than 0.08, also indicating an insufficient fit, but the SRMR is lower than 0.06, indicating a good fit. Thus, the model from the designometrics perspective cannot be fully accepted based on these criteria.

Regarding the factor loadings of the BUS-11, a visualisation of the BUS-11 factors structure from the designometrics perspective can be seen in the Figure 3. They vary from 0.96 (QCI-nINFO) to 1. Factors that consist of multiple items (ACF, QCF, QCI) have their loadings from 0.96 to 0.99, while single item factors (PS, TR) loadings are 1. All items explain at least 96% of the variance in each factor, and the total mean value of the variance in each factor explained is 98.5%.

Figure 3

Visualisation of the BUS-11 factors structure (designometrics perspective)



The correlations between the factors can be seen in Table 8. The correlation range is from -0.286 to 0.961, with the weakest correlation of -0.183 being between factors 1 (ACF) and 4 (PS) and the strongest correlation between factors 2 (QCF) and 3 (QCI). Besides factor 4 (PS) being negatively correlated with all other factors, although non-significant, all factors are strongly correlated with each other, and the mean value between all factors is 0.451 (0.902 without factor 4 (PS)).

Table 8

Correlation of the BUS-11 factors (designometrics perspective)

	ACF	QCF	QCI	PS	TR
ACF					
QCF	0.919**				
QCI	0.811**	0.961**			
PS	-0.183	-0.227	-0.210		
TR	0.903**	0.941**	0.878**	-0.286	

Note. *p<0.05, **p<0.01

Reliability

The internal consistency of BUS-11 was measured by Cronbach's alpha to see the level of reliability. Cronbach's alpha is greater than 0.7 from both the psychometrics perspective (α = .873) and the designometrics perspective (α = .972), indicating a good-to-great reliability based on the criteria set for reliability by Cortina (1993), and Taber (2017) described in Data Analysis.

Moreover, an inter-item correlation matrix for the BUS-11 items was created for both perspectives. From the psychometrics perspective (see Table 9), the correlations vary from -0.001 (between PVCY and TIME) to 0.738 (between ASST and ACC), with the mean inter-item correlation of 0.389. From the designometrics perspective (see Table 10), the correlations vary from -0.257 (also between PVCY and TIME) to 0.980 (between DET and FND), with the mean inter-item correlation of 0.705.

In addition, there are differences between the correlations of items that belong to the same factor and the correlations of items that do not belong to the same factor. From the psychometrics perspective (see Table 9), all items have a low correlation with factor 1 (ACF, items DET and FND) except the correlation between themselves (0.525), although only a mediocre correlation. To continue, all items have a low correlation with factor 4 (PS, item PVCY) and factor 5 (TR, item TIME), especially, the correlation between the factors 4 and 5 (-0.001), however, stronger correlations are visible between factors 2 (QCF, items COM, TRCK and DIG) and 3 (QCI, items ASST, aINFO, nINFO and ACC).

In comparison, from the designometrics perspective (see Table 10), the only low correlations, as well as negative, are with the factor 4 (PS, item PVCY), ranging from -0.102 to -0.257, whereas all other items are strongly correlated, ranging from 0.709 to 0.980.

Table 9

Correlation of the BUS-11 items (psychometrics perspective)

	DET	FND	COM	TRCK	DIG	ASST	aINFO	nINFO	ACC	PVCY	TIME
DET											
FND	0.525**										
COM	0.303**	0.227**									
TRCK	0.329**	0.182**	0.614**								
DIG	0.321**	0.219**	0.626**	0.521**							
ASST	0.292**	0.112*	0.659**	0.628**	0.566**						
aINFO	0.266**	0.136**	0.617**	0.612**	0.582**	0.705**					
nINFO	0.238**	0.138**	0.589**	0.604**	0.535**	0.617**	0.683**				
ACC	0.267**	0.158**	0.675**	0.637**	0.591**	0.738**	0.658**	0.632**			
PVCY	0.288**	0.178**	0.191**	0.256**	0.181**	0.248**	0.223**	0.256**	0.223**		
TIME	0.175**	0.224**	0.315**	0.300**	0.354**	0.317**	0.307**	0.234**	0.341**	-0.001	

Note. *p<0.05, **p<0.01; Red – correlation ≥ 0.5

Table 10

Correlation of the BUS-11 items (designometrics perspective)

	DET	FND	COM	TRCK	DIG	ASST	aINFO	nINFO	ACC	PVCY	TIME
DET											
FND	0.970**										
COM	0.935**	0.889**									
TRCK	0.904**	0.852**	0.971**								
DIG	0.930**	0.905**	0.982**	0.949**							
ASST	0.833**	0.773**	0.960**	0.963**	0.937**						
aINFO	0.789**	0.740**	0.927**	0.943**	0.912**	0.980**					
nINFO	0.794**	0.709**	0.907**	0.929**	0.861**	0.959**	0.953**				
ACC	0.880**	0.838**	0.974**	0.963**	0.966**	0.974**	0.958**	0.931**			
PVCY	-0.206	-0.102	-0.205	-0.225	-0.162	-0.185	-0.161	-0.175	-0.194		
TIME	0.904**	0.894**	0.932**	0.896**	0.957**	0.878**	0.848**	0.788**	0.937**	-0.257	

Note. *p<0.05, **p<0.01; Blue – correlation ≤ 0.5 (or between -0.5 and 0.5)

Validity

BUS-11 and UMUX-Lite

To check the validity of the BUS-11, the correlation between the BUS-11 and UMUX-Lite from the psychometrics perspective was performed using regression as part of the SEM. The results show a positive correlation between BUS-11 and UMUX-Lite (Estimate = 0.487, 95% CI [0.446, 0.529]). Thus, indicating that both BUS-11 and UMUX-Lite measure user satisfaction. The visualisation of the correlation can be seen in Figure 4. Strong correlations are with the factor 2 (QCF, -0.637) and the factor 3 (QCI, 0.789), however, none of the factors of BUS-11 are significantly related to UMUX-Lite (see Table 11).

Figure 4

Visualisation of the correlation between BUS-11 and UMUX-Lite

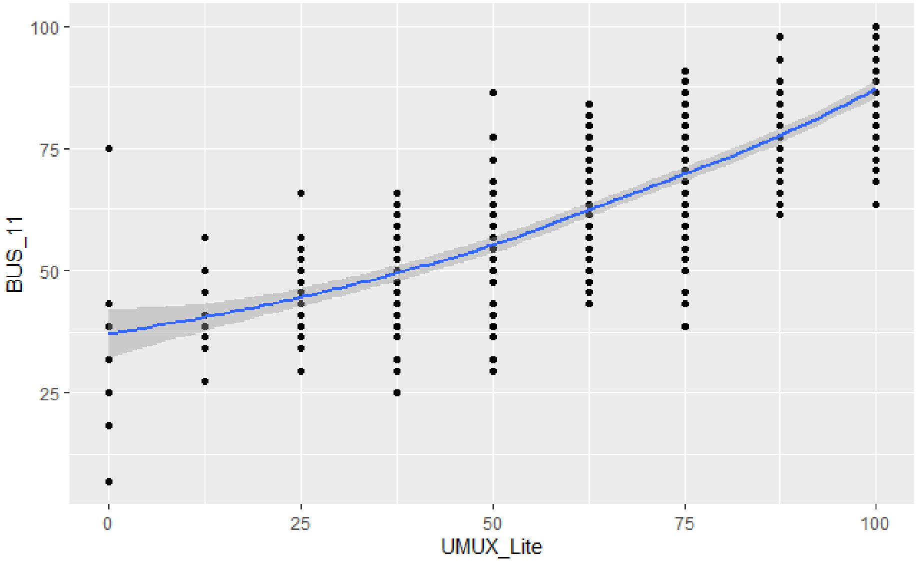


Table 11

Estimates between the five factors of the BUS-11 and UMUX-Lite

	ACF	QCF	QCI	PS	TR
UMUX-Lite	0.153	-0.637	0.789	-0.060	0.022

Note. *p<0.05, **p<0.01

RSME and Familiarity

BUS-11 and RSME

To see the correlation between the BUS-11 and RSME from the psychometrics perspective, regression as part of the SEM was performed. The results show a negative correlation between BUS-11 and RSME (Estimate = -0.106, 95% CI [-0.153, -0.059]). The visualisation of the correlation can be seen in Figure 5. There are three negative correlations with factors 1 (ACF), 3 (QCI), and 5 (TR), and two positive correlations with factors 2 (QCF) and 4 (PS), however, only factors 1 (ACF) and 3 (QCI) of BUS-11 are significantly related to RSME (see Table 12).

Figure 5

Visualisation of the correlation between BUS-11 and RSME

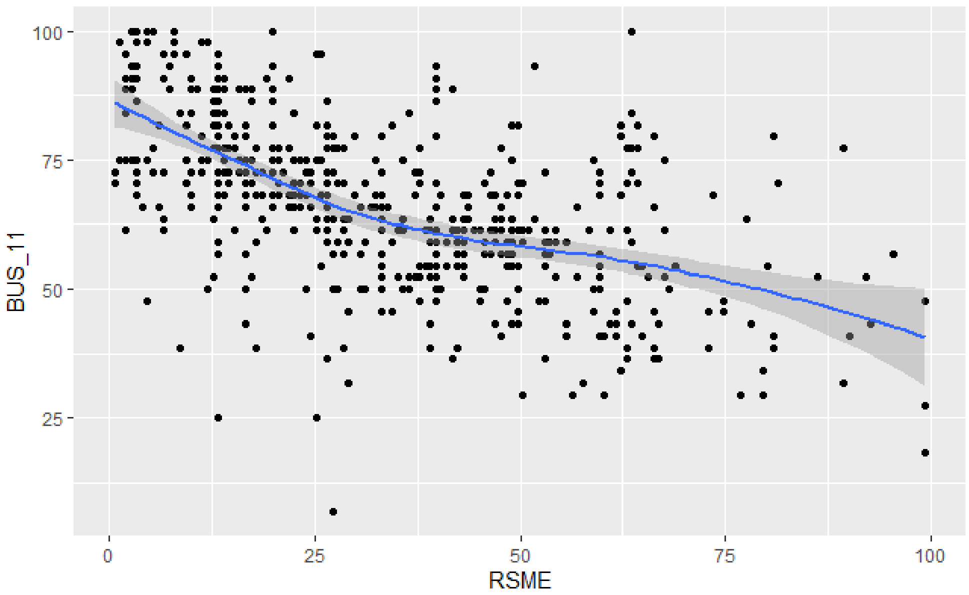


Table 12

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

	ACF	QCF	QCI	PS	TR
RSME	-0.090*	0.384	-0.473*	0.022	-0.016

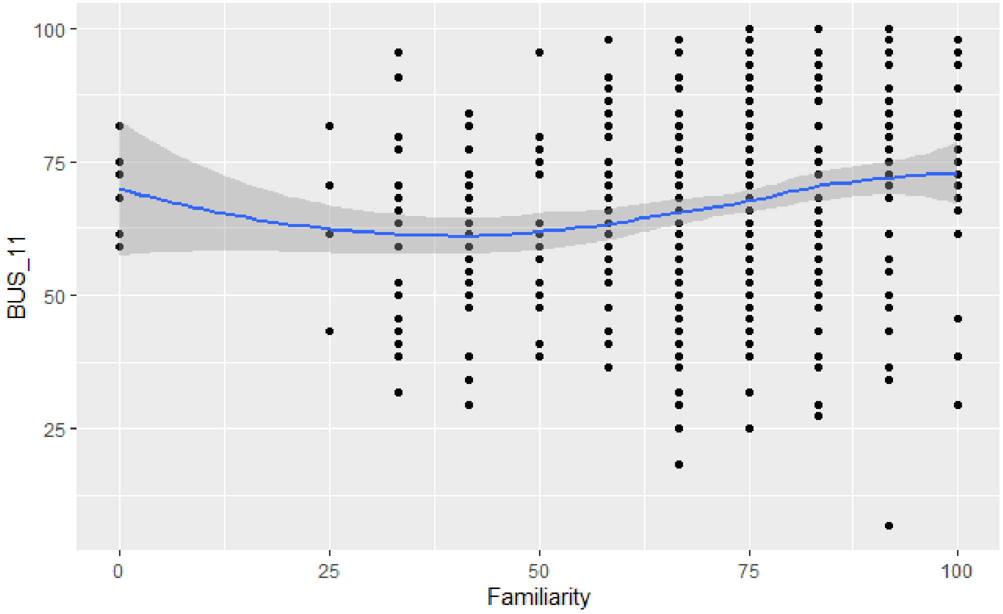
Note. *p<0.05, **p<0.01

BUS-11 and Familiarity

To see the correlation between the BUS-11 and familiarity from the psychometrics perspective, a regression as part of the SEM was performed. The results show a positive correlation between BUS-11 and familiarity, although a weak one, as it is close to zero (Estimate = 0.069, 95% CI [0.021, 0.118]). The visualisation of the correlation with the smooth curve (LOESS) can be seen in Figure 6.

Figure 6

Visualisation of the correlation between BUS-11 and Familiarity



Discussion

The general aim of this study was to advance previous work done on the new scale BUS-11 to assess satisfaction with chatbots, with a focus on the assessment of chatbots to further streamline the reliability and validity of the scale. For this, the five-factor structure of the BUS-11 was re-tested using other chatbots (in Dutch and/or English), the reliability of the scale was assessed, its internal validity was assessed by performing confirmatory factor analysis, and its external validity by using the UMUX-Lite scale. The data were explored from both psychometrics and designometrics perspectives.

Regarding the first research question ("Can the five-factor structure of the BUS-11 be confirmed from the psychometrics and designometrics perspectives?"), from the psychometrics perspective, the results show a good fit of the five-structure model, and that the factorial structure is in line with the previous work by Borsci et al. (2022b). The factor loadings are rather high (mean value of 0.823), and the correlations between the factors are rather weak with the mean value of 0.318. However, there is a strong correlation (0.823) between the factor 2 (QCF) and 3 (QCI). This could be due to the potential similarities between those two factors, measuring similar constructs, as the factor 2 is "Perceived quality of the chatbot" and the factor 3 is "Perceived quality of conversation and information provided" (see Appendix D). Moreover, responding to rating scales can lead to an answer that is not fully representative. Meaning, the participants had seen a chatbot only in a particular content of that chatbot, so it is difficult to see the distinction between the factors from their perspective, as chatbots may work and be perceived differently in a different context. Overall, the results indicate that there are five different factors.

From the designometrics perspective, out of all four criteria for an acceptable model (CFI, TLI, RMSEA, SRMR), only SRMR is within the requirements, hence, showing an insufficient fit of the five-structure model. This could be due to the differences in the factorial structure of the BUS-11 between those two perspectives and/or the fact that the sample size of the designometrics perspective is significantly smaller compared to the psychometrics perspective. Regarding the factor loadings, they are greater than from the psychometrics perspective (mean value of 0.985), and the correlations between the factors, except with the factor 4 (PS, low and negative), are strong (≥0.811). This indicates that four out of five factors measure the same construct and that they could be combined, resulting in a model with fewer factors that could be a better fit for the designometrics perspective.

Based on the results, essentially, the BUS-11 scale can measure user satisfaction of chatbots from the psychometrics perspective but cannot be fully confirmed from the designomentrics perspective. Nevertheless, the sample size of the chatbots for the desigonmetrics perspective should be considered, as a larger sample size would give more accurate results.

The second research question was “Are the items of the BUS-11 reliable?”. The internal consistency (reliability) of the BUS-11 from both the psychometrics and designometrics perspective is high (α = .873 and α = .972, respectively). Hence, the items of the BUS-11 are reliable from both perspectives.

To continue, from the psychometrics perspective, the correlations of items vary from -0.001 to 0.738, with the mean inter-item correlation of 0.389, and from the designometrics perspective, the correlations vary from -0.257 to 0.980, with a mean inter-item correlation of 0.705. Although the range is broader from the designometrics perspective because of the factor 4 (PS) being negative, the mean-inter-item correlation is significantly greater than from the psychometrics perspective, which is also in line with the findings regarding the first research question.

Additionally, there are differences between the correlations of items that belong to the same factor and the correlations of items that do not belong to the same factor. From the psychometrics perspective, all items have low correlation with the factor 1 (ACF) items, except the correlation between themselves, although only a mediocre correlation (0.525). To continue, all items have low correlation with the factor 4 (PS) and the factor 5 (TR), especially, the correlation between the factors 4 and 5 (-0.001), however, stronger correlations are visible amongst the items of the factors 2 (QCF) and 3 (QCI). These strong correlations are simply because these items belong to the same factor, therefore, also being more related to each other, as well as that factors 2 and 3 have potential similarities, as mentioned earlier.

On the other hand, from the designometrics perspective, the only low correlations, as well as negative, are with the factor 4 (PS), ranging from -0.102 to -0.257, whereas all other items are strongly correlated, ranging from 0.709 to 0.980. Once again, like correlations between factors before, this could be due to the differences in the factorial structure of the BUS-11 between those two perspectives, and, as there are strong correlations between factors, there are also strong correlations between items, indicating a similar construct being measured.

The third research question was “Do the results of the BUS-11 correlate positively with the results of the UMUX-Lite?”, which is about the validity of the BUS-11. There is a moderate positive correlation between BUS-11 and UMUX-Lite (0.487), as there also was a positive significant relationship between those two scales in the study by Borsci et al. (2022b), indicating that both scales measure a similar construct. UMUX-Lite strongly correlates with the factor 2 (QCF), although negatively (-0.637), and the strongest correlation is with the factor 3 (QCI, 0.789). Factor 3 could be compared to the first item of the UMUX-Lite – “The options offered by the chatbot meet my requirements.” –, meaning, they measure user satisfaction in a similar manner. The weak correlations other factors could be explained that different aspects of user satisfaction are measured, therefore, potentially indicating a broader measurement of user satisfaction by the BUS-11 compared to the UMUX-Lite.

Thus, it can be said that BUS-11 is a valid measurement tool for user satisfaction of chatbots. By confirming and, potentially, improving this scale even more, there will, finally, be a scale for user satisfaction in the context of chatbots, which would be beneficial with the growth of AI and chatbots.

The final research question was “How does mental effort and familiarity (previous experience) of chatbots affect the user satisfaction?”. Regarding the mental effort, the results show a negative correlation between BUS-11 and RSME (-0.106), however, a weak one. There are three negative correlations with factors 1 (QCF), 3 (QCI) and 5 (TR), and two positive correlations with factors 2 (QCF) and 4 (PS). Nevertheless, this indicates that the lower the mental effort put, the higher the user satisfaction, and, therefore, during the development of chatbots, designers should consider mental (cognitive) workload by optimising the design of chatbots, in order to make users as satisfied with those chatbots as possible. Although the correlation is understandable and expected, more research should be done to fully confirm this.

In terms of the effect of familiarity (previous experience), the results show a positive correlation between BUS-11 and familiarity. Although, as it is a weak one and close to zero (0.069), it can be said that, basically, there is no relationship between the familiarity and the user satisfaction. Based on this, these findings are not in line with previous work by Borsci et al. (2015) and McLellan et al. (2012). However, it is worth noting, that this could be affected by the unstandardised way of measuring familiarity in this study. In addition, the mean age of participants was 22.3, and their level of English varied from medium to high, as well as most of the participants, regarding the frequency of chatbot usage, reported “never” or “rarely”. Meaning, younger generations tend to be more familiar with technology itself, and with sufficient knowledge of English, as in this case, this nullifies the effect of familiarity on satisfaction, however, only potentially, and this should be researched more, as the data from various studies are inconsistent.

Limitations and recommendations

For this study, three limitations can be mentioned. The first one, and perhaps the more important one, is the sample size of the chatbots for the designometrics perspective, as a larger sample size would give more accurate results. It could give a further insight also about the potentially different factorial structure of the BUS-11 from the designometrics perspective, as the results suggested. Even though there was an opportunity to combine the gathered data from Huijsmans and Borsci (2022) and Braun and Borsci (2023) studies to mediate this limitation, the sample size was still small (31), which could be a reason for the obtained findings. Therefore, the results from the designometrics perspective should be interpreted accordingly. Noteworthy, a larger sample of chatbots comes with difficulties, as participants would have to invest more time to interact with more chatbots in one go, which could lead to less accurate results due to loss of concentration and motivation, for example. However, the larger sample could also be obtained by increasing the database of chatbots, from which the randomised selection of chatbots is made, as well as increasing the amount of people participating in these studies. Moreover, the data from multiple studies can be combined to alter the sample size, as it was done in this study, and the results from these combined studies could be used in the future studies.

The second limitation is the measurement of familiarity (previous experience) of chatbots. It was done in an unstandardised way. Consequently, the relationship between familiarity and user satisfaction cannot be established and concluded. So, in a similar manner as with the designometrics perspective, the results of the effects of familiarity on user satisfaction should be interpreted accordingly. Therefore, more research should be done, as the data from various studies are inconsistent, as well as getting participants with more variety in terms of their familiarity with chatbots or technology itself could provide additional data towards the relationship. Additionally, a better, more standardised method to measure familiarity of chatbots or perhaps of technology itself should be considered.

The last limitation is regarding the included factors in this study, such as familiarity. Only familiarity was included in this study as it could be connected to other factors briefly mentioned in this study – productivity, humanisation of chatbots, creepiness, and trust. Nevertheless, these other factors can have an effect on user experience and user satisfaction individually, so it would be beneficial to explore their effects as, for instance, did Braun and Borsci (2023) about trust and Huijsmans and Borsci (2022) on another factor – age.

As another general recommendation, more time and examination could be given towards the selection of chatbots and their task/scenario descriptions. Chatbots differ in many ways, such as complexity, and, even though the descriptions were tried to be tailored to the chatbots in the best way possible whilst keeping all of them concise, the gathered data showed that some chatbots would have required, for example, more elaboration. Some chatbots and tasks were easier than others and, apparently, some tasks could not be completed because the chatbot did not appear or was not found, although the chatbot selection was made in such a way that this should not have occurred. In overall, more attention should be given in creating the list of chatbots and the tasks in order to make it more balanced and, potentially, obtaining more accurate results with all the scales involved in a study.

Conclusion

The present study contributed towards insights about the factorial structure of the BUS-11 from both psychometrics and designometrics perspectives. The five-factor structure model was confirmed from the psychometrics perspective, however, more research should be done from the designometrics perspective, as the results suggested a different (smaller) factorial structure that could be a better fit for this perspective. The results from this study provided deeper insights of the BUS-11 items and factors, and their relations to adjust and improve the scale even more.

The items of the BUS-11 proved to be reliable from both perspectives, and the analysis of the internal consistency (reliability) and validity of the BUS-11 scale show that BUS-11 is a valid measurement tool for user satisfaction of chatbots. By confirming and, potentially, improving this scale even more, there will, finally, be a scale for user satisfaction in the context of chatbots, which would be beneficial with the growth of AI and chatbots.

The results showed a negative correlation between BUS-11 and RSME, however, a weak one. Nevertheless, this indicates that the lower the mental effort put, the higher the user satisfaction, and, therefore, during the development of chatbots, designers should consider mental (cognitive) workload by optimising the design of chatbots, in order to make users as satisfied with those chatbots as possible.

Lastly, the results showed a positive correlation between BUS-11 and familiarity. Although, as it is a weak one and close to zero, it can be said that there is no relationship between the familiarity and the user satisfaction based on this study.

References list

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based Chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7

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

Alimohammadi, I., Zabiholah, D., Rahmani, N., Parsazadeh, B. & Yeganeh, R. (2019). Validity and reliability of rating scale mental effort, integrated workload scale, and overall workload scale in Iran. *International Journal of Occupational Hygiene*, 301-304.

Angga, P. A., Fachri, W. E., Elevanita, A., Suryadi, & Agushinta, R. D. (2015). Design of chatbot with 3D Avatar, voice interface, and facial expression. *2015 International Conference on Science in Information Technology (ICSITech)*. https://doi.org/10.1109/icsitech.2015.7407826

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

Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. (2020). I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-Powered Service Agents. *Telematics and Informatics*, *54*, 101473. https://doi.org/10.1016/j.tele.2020.101473

Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., Antunes, R., Gomes, M., Gatti, C., Vanzin, M., Junior, S. C., Silva, E., & Moreira, C. (2020). Conversational agents in Business: A Systematic Literature Review and Future

Research Directions. *Computer Science Review, 36*, 100239. https://doi.org/10.1016/j.cosrev.2020.100239

Balaji, D., & Borsci, S. (2019). *Assessing user satisfaction with information chatbots: A preliminary investigation*. (Master thesis). University of Twente, Enschede, Netherlands.

Bekmanis, N., & Borsci, S. (2020). *Do you trust technology? An exploratory investigation on the concept of trustworthiness and the role of memory*. (Bachelor thesis). University of Twente, Enschede, The Netherlands.

Ben Mimoun, M. S., Poncin, I., & Garnier, M. (2017). Animated conversational agents and e consumer productivity: The roles of agents and individual characteristics. *Information & Management*, *54*(5), 545–559. https://doi.org/10.1016/j.im.2016.11.008

Bevan, N. (1995). Usability is quality of use. *Advances in Human Factors/Ergonomics*, 349 354. https://doi.org/10.1016/s0921-2647(06)80241-8

Braun, M., & Borsci, S. (2023). *Evaluating the chatbot usability scale: a psychometric and designometric perspective*. (Master thesis). University of Twente, Enschede, The Netherlands.

Brooke, J. (1996). Sus: A 'quick and dirty' usability scale. *Usability Evaluation In Industry*, 207–212. https://doi.org/10.1201/9781498710411-35

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. (2021). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and*

Ubiquitous Computing, *26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

Borsci, S., Prati, E., Federici, S., Malizia, A., Schmettow, M., & Chamberlain, A. (2022). “Ciao AI”: The Italian adaptation and validation of the chatbot usability scale. https://doi.org/10.31234/osf.io/3hcgy

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

CGS (2019). *CGS Survey reveals consumers prefer a hybrid AI/Human approach to customer service. Is there chatbot fatigue?* https://www. cgsinc.com/en/resources/2019-CGS-Customer-Service-Chatbots- Channels-Survey

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

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionary of statistics*. Cambridge, UK: Cambridge University Press. Retrieved from http://www.stewartschultz.com/statistics/books/Cambridge%20Dictionary%20Statistcs%204th.pdf

Federici, S., de Filippis, M. L., Mele, M. L., Borsci, S., Bracalenti, M., Gaudino, G., Cocco, A., Amendola, M., & Simonetti, E. (2020). Inside pandora’s box: A systematic review of the assessment of the perceived quality of Chatbots for people with disabilities or special needs. *Disability and Rehabilitation: Assistive Technology*, *15*(7), 832–837. https://doi.org/10.1080/17483107.2020.1775313

Fernandes, A. (2018, November 9). *NLP, NLU, NLG and how Chatbots Work*. Medium. Retrieved from https://chatbotslife.com/nlp-nlu-nlg-and-how-chatbots-work dd7861dfc9df

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., & Brandtzæg, P. B. (2017). Chatbots and the New World of HCI. *Interactions*, *24*(4), 38–42. https://doi.org/10.1145/3085558

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2019). Different chatbots for different purposes: Towards a typology of Chatbots to understand interaction design. *Internet Science*, 145–156. https://doi.org/10.1007/978-3-030-17705-8_13

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega*, *28*(6), 725–737. https://doi.org/10.1016/s0305-0483(00)00021-9

Gnewuch, U., Morana, S., Adam, M. T. P., and Maedche, A. (2018). Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. *Proceedings of the 26th European Conference on Information Systems* (ECIS), Portsmouth, United Kingdom, June 23-28.

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

Hoedemaeker, M. (2002). *Summary Description of Workload Indicators: WP1 Workload Measures. Human Machine Interface and the Safety of Traffic in Europe Growth Project.* GRD1-2000-25361. HASTE. Institute for Transport Studies. Leeds, UK: University of Leeds.

Hopper, E. (2021). *An Introduction to Rogerian Therapy*. Retrieved from https://www.thoughtco.com/rogerian-therapy-4171932

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly, 37*(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391

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

Huijsmans, M., & Borsci, S. (2022). *The chatbot usability scale: an evaluation of the Dutch version of the BUS-11*. (Master thesis). University of Twente, Enschede, The Netherlands.

Inman, J. J., & Nikolova, H. (2017). Shopper-facing retail technology: A retailer adoption decision framework incorporating shopper attitudes and privacy concerns. *Journal of Retailing, 93*(1), 7–28. https://doi.org/10.1016/j.jretai.2016.12.006

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

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2020). Privacy concerns in chatbot interactions. *Chatbot Research and Design*, 34–48. https://doi.org/10.1007/978-3-030-39540-7_3

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

Jenkins, M.-C., Churchill, R., Cox, S., & Smith, D. (2007). Analysis of user interaction with service oriented Chatbot Systems. *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, 76–83. https://doi.org/10.1007/978-3-540 73110-8_9

Khalid, H. M., Helander, M. G., & Lin, M.-H. (2021). Determinants of trust in human-robot interaction: Modeling, measuring, and predicting. *Trust in Human-Robot Interaction*, 85–121. https://doi.org/10.1016/b978-0-12-819472-0.00004-6

Khalid, H. M., Shiung, L. W., Sheng, V. B., & Helander, M. G. (2018). Trust of Virtual Agent in Multi Actor Interactions. *Journal of Robotics, Networking and Artificial Life*, *4*(4), 295. https://doi.org/10.2991/jrnal.2018.4.4.8

Kim, Y., & Peterson, R. A. (2017). A meta-analysis of online trust relationships in e commerce. *Journal of Interactive Marketing*, *38*, 44–54. https://doi.org/10.1016/j.intmar.2017.01.001

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of Bots. *Proceedings of the 2017 Conference on Designing Interactive Systems*. https://doi.org/10.1145/3064663.3064672

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 Annual Meeting*, *57*(1), 197–201. https://doi.org/10.1177/1541931213571044

Langer, M., & König, C. J. (2018). Introducing and testing the creepiness of situation scale (cross). *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02220

Lee, H.-J. (2017). Personality determinants of need for interaction with a retail employee and its impact on self-service technology (SST) usage intentions. *Journal of Research in Interactive Marketing*, *11*(3), 214–231. https://doi.org/10.1108/jrim-04-2016-0036

Lewis, J. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*(3), 463 488. https://doi.org/10.1207/s15327590ijhc143&4_11

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, Sus, and umux. *International*

Journal of Human–Computer Interaction, *34*(12), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). Usability metric for user experience--lite. *PsycTESTS Dataset*. https://doi.org/10.1037/t81953-000

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The sus, umux-lite, and Altusability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505. https://doi.org/10.1080/10447318.2015.1064654

Lewis, J., & Sauro, J. (2020). *How to convert between five- and seven-point scales*. MeasuringU. Retrieved from https://measuringu.com/convert-point-scales/

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with Computers*, *15*(3), 429–452. https://doi.org/10.1016/s0953 5438(02)00063-2

Ling, E. C., Tussyadiah, I., Tuomi, A., Stienmetz, J., & Ioannou, A. (2021). Factors influencing users' adoption and use of conversational agents: A systematic review. *Psychology & Marketing*, *38*(7), 1031–1051. https://doi.org/10.1002/mar.21491

Luhmann, N. (1979). *Trust and power*. Wiley.

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*. https://doi.org/10.1287/mksc.2019.1192

Lutz, C., Hoffmann, C. P., & Ranzini, G. (2020). Data capitalism and the user: An exploration of privacy cynicism in Germany. *New Media & Society*, *22*(7), 1168 1187. https://doi.org/10.1177/1461444820912544

Mani, Z., & Chouk, I. (2016). Drivers of consumers’ resistance to smart products. *Journal of Marketing Management*, *33*(1-2), 76–97. https://doi.org/10.1080/0267257x.2016.1245212

Mauldin, M. L. (1994). ChatterBots, tinyMuds, and the turing test entering the loebner prize competition. *Proceedings of the National Conference on Artificial Intelligence*, *1*, 16 21.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in new organizational relationships. *Academy of Management Review*, *23*(3), 473–490. https://doi.org/10.5465/amr.1998.926622

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on System Usability Scale ratings. *Journal of Usability Studies*, *7*, 56–67.

Molnár, G., & Zoltán, S. (2018). The role of Chatbots in formal education. *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*. https://doi.org/10.1109/sisy.2018.8524609

Mozafari, N., Weiger, W. H., & Hammerschmidt, M. (2021). Trust me, I'm a bot – repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management*, *33*(2), 221–245. https://doi.org/10.1108/josm-10-2020-0380

Paikari, E., & van der Hoek, A. (2018). A framework for understanding Chatbots and their future. *Proceedings of the 11th International Workshop on Cooperative and Human Aspects of Software Engineering*. https://doi.org/10.1145/3195836.3195859

Park, J. K., Ahn, J., Thavisay, T., & Ren, T. (2019). Examining the role of anxiety and social influence in multi-benefits of Mobile Payment Service. *Journal of Retailing and Consumer Services*, *47*, 140–149. https://doi.org/10.1016/j.jretconser.2018.11.015

Pollmann, N., & Borsci, S. (2021). *Testing a usability scale for chatbots: The effect of familiarity on satisfaction ratings*. (Bachelor thesis). University of Twente, Enschede, The Netherlands.

Powton, M. (2018, June 4). *A visual history of chatbots*. Medium. Retrieved from https://chatbotsmagazine.com/a-visual-history-of-chatbots-8bf3b31dbfb2

Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A Social Relationship Perspective to designing information systems. *Journal of Management Information Systems*, *25*(4), 145–182. https://doi.org/10.2753/mis0742 1222250405

Rajaobelina, L., Prom Tep, S., Arcand, M., & Ricard, L. (2021). Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychology and Marketing*, *38*(12), 2339–2356. https://doi.org/10.1002/MAR.21548

Ranoliya, B. R., Raghuwanshi, N., & Singh, S. (2017). Chatbot for university related faqs. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. https://doi.org/10.1109/icacci.2017.8126057

Rosseel, Y, Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., … Du, H. (2021). lavaan: Latent variable analysis (Version 0.6-8). Retrieved from https://cran.r-project.org/web//packages/lavaan/lavaan.pdf

R Core Team. (2022). R: A language and environment for statistical computing (Version 4.2.3).

Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R- project.org

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/1978942.1979266

Schmettow, M. (2020). Psychometrics and design-o-metric models. *In New statistics for Design Researchers: A Bayesian workflow in Tidy R*. Springer Nature. Retrieved from https://schmettow.github.io/New_Stats/

Schmettow, M., & Borsci, S. (2020). RPubs - It´s a cuboid, stupid! On the designometric perspective and the psychometric fallacy. Retrieved from https://rpubs.com/schmettow/Designometrics_1

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

Scopus. (n.d.). Retrieved from https://www.scopus.com/home.uri

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? comparison of eliza with modern dialogue systems. *Computers in Human Behavior, 58*, 278–295. https://doi.org/10.1016/j.chb.2016.01.004

Taber, K. S. (2017). 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

Tariverdiyeva, G., & Borsci, S. (2019). *Chatbots' perceived usability in information retrieval tasks: An exploratory analysis*. (Master thesis). University of Twente, Enschede, The Netherlands.

Tene, O., & Polonetsky, J. (2014). A theory of creepy: Technology, privacy, and shifting social norms. *Yale Journal of Law and Technology*, 16,59–102.

Thomaz, F., Salge, C., Karahanna, E., & Hulland, J. (2019). Learning from the dark web: Leveraging conversational agents in the era of hyper-privacy to enhance marketing. *Journal of the Academy of Marketing Science, 48*(1), 43–63. https://doi.org/10.1007/s11747-019-00704-3

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in Mental Health: A review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry, 64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on Social Media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025496

Zamora, J. (2017). I'm sorry, Dave, I'm afraid I can't do that. *Proceedings of the 5th International Conference on Human Agent Interaction*. https://doi.org/10.1145/3125739.3125766

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

Appendix

Appendix A – Informed Consent

Taking part in the study

I have read and understood the study information. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason. I understand that taking part in the study involves me interacting with different chatbots. The whole experiment will take about 40 minutes. I understand that for participating in the study there are no known risks involved. I am at least 18 years old.

Use of the information in the study

I understand that taking part in the study involves answering questions about my demographics, performing tasks, and interacting with chatbots online, and answering questions about each of the chatbots I have interacted with online.

Future use and reuse of the information by others

I understand that information I provide will be used for a master thesis. I understand that before the information is achieved it will be anonymized by removing name and other information that could track me back. I give permission for the answers that I provide in this survey to be archived in a safe data repository so it can be used for future research and learning.

Contact Information for Questions about Your Rights as a Research Participant

If you ever have any questions at any time before, during or after this session, you can email us: **n.bekmanis@student.utwente.nl** and our supervisor can be reached at **s.borsci@utwente.nl**. If you have questions about your rights as a research participant or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by **ethicscommittee-bms@utwente.nl**.

Yes, I understand the text above and I agree to participate in this study.

No, I do not agree, and I want to end this session.

Appendix B – The first part of this study

What is your age?

What is your gender (as assigned at birth)?

○ Male

○ Female

What is your nationality?

○ Dutch

○ German

○ Latvian

○ Other (please, specify)

What is your level of English?

○ Beginner (A1)

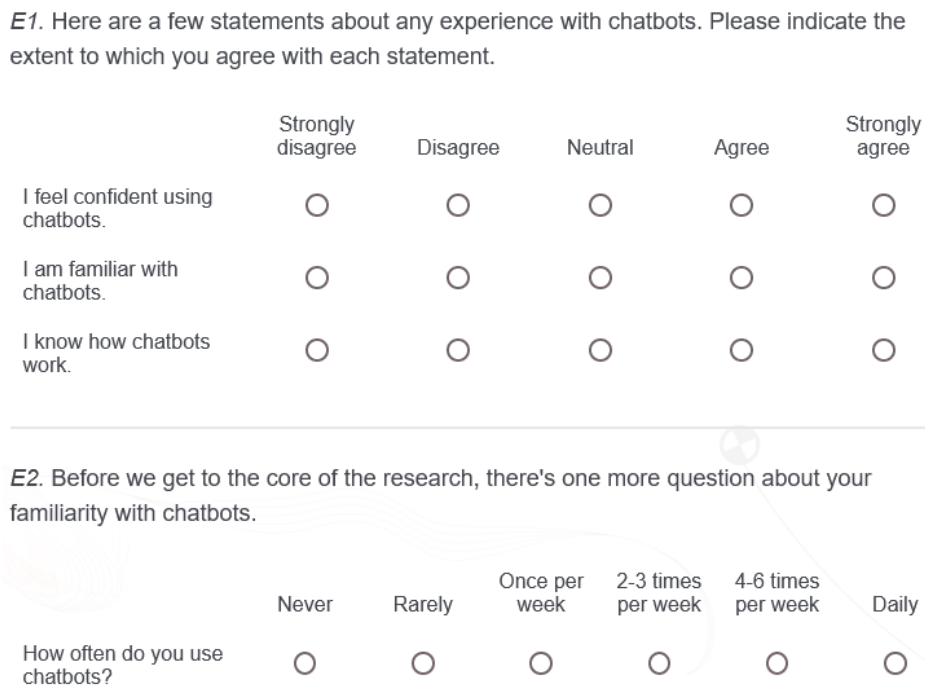
○ Elementary (A2)

○ Intermediate (B1)

○ Upper intermediate (B2)

○ Advanced (C1)

○ Proficient (C2)



E1. Here are a few statements about any experience with chatbots. Please indicate the extent to which you agree with each statement.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I feel confident using chatbots.	○	○	○	○	○
I am familiar with chatbots.	○	○	○	○	○
I know how chatbots work.	○	○	○	○	○

E2. Before we get to the core of the research, there's one more question about your familiarity with chatbots.

	Never	Rarely	Once per week	2-3 times per week	4-6 times per week	Daily
How often do you use chatbots?	○	○	○	○	○	○

Explanation

For each chatbot, you will be given a scenario, tasks and a link to a website with that chatbot. You are supposed to find the chatbot on the website and start communicating with it. When you have completed tasks given in a scenario, come back to the survey and complete the questions about that chatbot. In total, you will communicate with 6 different chatbots.

Note: a chatbot can redirect you to a page with information. If this page contains the information you need, you have completed the task. You never have to fill in personal information.

We are interested in the quality of chatbots and the user experience with them. Try to accomplish your tasks, but you do not have to reach them if it seems impossible, for example, without involving a live agent or filling in a contact form. When you are done with the interaction (successful or not), please, return to the survey.

For the best experience, please, turn off any kind of (ad) blockers on your internet browser.

**in this survey, can also be referred to a virtual/digital assistant and/or a name of that chatbot*

Appendix C – Chatbots, scenarios/tasks, links

WestJet – https://www.westjet.com/en-ca

You want to go to Berlin (Germany), and you want to know how you can get there. Also, in case something changes for you, you are interested in their policy for refunds. Consult WestJet's digital travel assistant Juliet for this information.

Singtel – https://www.singtel.com/

You are looking for a new phone and you are interested in iPhone 13. Consult Singtel's virtual assistant Shirley to find out:

- How can you get one?
- Its promotions
- Different mobile plans
- Delivery possibilities

Drift – https://www.drift.com/

You have a company with less than 50 employees, and you are interested in how Drift can help you accelerate revenue. In addition, you want to know, if there are any risks with Drift, as well as how can you apply for Drift. Consult DriftBot for this information.

NatWest – https://www.natwest.com/

Your online transaction with your debit card for your existing mortgage has been declined, and you want to find out the possible reasons. Consult Cora, the digital assistant, for this information.

SeattleBallooning – https://seattleballooning.com/

You are planning to go for a balloon flight in June with your family. You have children, so you are wondering if there are any age restrictions. Your family worries about the weather, so you want to know what happens, if it starts to rain. Also, you are interested in the duration and the prices of the flights. Consult Seattle Balloon Assistant for this information.

Amtrak – https://www.amtrak.com/home.html

You are planning a long trip by train. You want to know what the sleeping and dining options on a train are, and if trains are pet friendly. Additionally, during these times, you want to know if there are any covid safety measures. Consult the virtual assistant Julie for this information.

Bupa – https://www.bupa.co.uk/

You are feeling unwell, and you want to find out if and how it is possible to test for COVID-19. Additionally, you want to test for detectable COVID-19 antibodies, and you were wondering how and when you can contact them by phone. Seek the support from the Bupa Virtual Assistant for this information.

Zoom – https://zoom.us/

You want to have group meetings with up to 50 people for hours and some storage for recordings. Your budget is 200 Euros per year. Consult Bolt, Zoom's Virtual Assistant, to find out which plan suits your needs the best.

MailChimp – https://mailchimp.com/

You want to boost your business in communication with your clients, so you are interested in how Mailchimp could help you with that, as you want to create automated e-mails etc. You also want to know how you can create classic automations. In addition, try to explore marketing plans and their pricing. Consult Mailchimp Assistant for this information.

Appendix D – BUS-11

Factor	Item
1 - Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable (e.g., the possibility to modify the settings of the chatbot, make the avatar visible or not etc.).
	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.

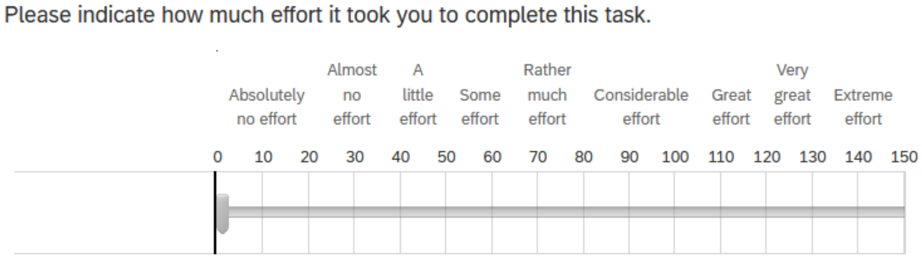
The old description of the item one used in Huijsmans and Borsci (2022) and Braun and Borsci (2023) studies – "The chatbot function was easily detectable."

Appendix E – Usability Metric for User Experience Lite (UMUX-Lite)

1 – The options offered by the chatbot meet my requirements.

2 – The chatbot is easy to use.

Appendix F – Rating Scale Mental Effort (RSME)



Please indicate how much effort it took you to complete this task.

Appendix G – The code used for the analysis in R

#Setup

##Libraries

```
## tidyverse
library(tidyverse)
## data manipulation
library(openxlsx)
library(readxl)
library(dplyr)
library(polynom)
## plotting
library(gridExtra)
library(ggplot2)
library(corrplot)

## corrplot 0.92 loaded

library(ggpubr)
## regression models
library(rstanarm)
library(car)
library(brms)
## CFA and SEM (+ plotting)
library(lavaan)
library(lavaanPlot)
library(semPlot)
## utilities for computing indices of model quality and goodness of fit
library(performance)
## reliability
library(psy)
library(psych)
## other
library(Hmisc)
library(devtools) ## only needed for installing from Github
library(knitr)
library(printr)
library(rmarkdown)
## non-CRAN packages

#install_github("schmettow/mascutils")
#install_github("schmettow/bayr")

## various utility functions for math, data manipulation, simulation and reporting
library(mascutils)
## for unified reporting of Bayesian regression results
library(bayr)
```

##Import, read and view

```
#importing, reading and viewing the data
PsyMx_BUS <- read_excel("PSYMX_BUS.xlsx")
DesMx_combined_BUS <- read_excel("DESMX_MB+MH+NB_BUS.xlsx")

#view(PsyMx_BUS)
#view(DesMx_combined_BUS)
```

#Internal consistency

##Cronbach's Alpha

###Psychometrics perspective

```
cronbach(PsyMx_BUS)
```

###Designometrics perspective

```
cronbach(DesMx_combined_BUS)
```

#Confirmatory Factor Analysis

##Psychometrics perspective

###Defining and fitting the model ####CFA

```
#Factors
#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

M_PsyMx <- 'ACF =~ DET+ FND
        QCF =~ COM + DIG + TRCK
        QCI =~ ASST + aINFO + nINFO + ACC
        PS =~ PVCY
        TR =~ TIME'

PsyFit_PsyMx <- cfa(M_PsyMx, data = PsyMx_BUS, std.lv = TRUE)

summary(PsyFit_PsyMx, standardized = TRUE, ci = TRUE, fit.measures = TRUE, rsq = TRUE)

semPaths(PsyFit_PsyMx,whatLabels = "std",edge.label.cex = 1, style = "lisrel", residScale = 8, layout = "tree3", theme = "colorblind", rotation = 2, what = "std", nChartNodes = 0, curvePivot = TRUE, sizeMan = 6, sizeLat = 12)
```

####SEM+Regressions

```
PsyMx_BUS_SEM <- read_excel("PSYMX_BUS_SEM.xlsx")

#Factors
#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

M_PsyMx_SEM <- 'F1 =~ I1 + I2
        F2 =~ I3 + I4 + I5
        F3 =~ I6 + I7 + I8 + I9
        F4 =~ I10
        F5 =~ I11

BUS ~ UMUX_L + RSME + Fam + Fre + Eng

UMUX_L ~ F1 + F2 +F3 + F4 + F5
RSME ~ F1 + F2 +F3 + F4 + F5'

PsyFit_PsyMx_SEM <- sem(M_PsyMx_SEM, data = PsyMx_BUS_SEM, std.lv = TRUE)

summary(PsyFit_PsyMx_SEM, standardized = TRUE, ci = TRUE, fit.measures = TRUE, rsq
= TRUE)
```

###Inter item correlation

```
item_intercor(PsyMx_BUS)

cor(PsyMx_BUS, method = "pearson")

rcorr(as.matrix(PsyMx_BUS),type = "pearson")
```

###Factors correlation

```
PsyMx_F <- read_excel("PSYMX_F.xlsx")

item_intercor(PsyMx_F)

cor(PsyMx_F, method = "pearson")

rcorr(as.matrix(PsyMx_F),type = "pearson")
```

##Designometrics perspective

###Defining and fitting the model ####CFA

```
#Factors name
#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

M_DesMx <- 'ACF =~ DET+ FND
        QCF =~ COM + DIG + TRCK
        QCI =~ ASST + aINFO + nINFO + ACC
        PS =~ PVCY
        TR =~ TIME'

DesFit_DesMx <- cfa(M_DesMx, data = DesMx_combined_BUS, std.lv = TRUE)

summary(DesFit_DesMx, standardized = TRUE, ci = TRUE, fit.measures = TRUE, rsq = TRUE)

semPaths(DesFit_DesMx,whatLabels = "std",edge.label.cex = 1, style = "lisrel", residScale = 8, layout = "tree3", theme = "colorblind", rotation = 2, what = "std", nChartNodes = 0, curvePivot = TRUE, sizeMan = 6, sizeLat = 12)
```

###Inter item correlation

```
item_intercor(DesMx_combined_BUS)

cor(DesMx_combined_BUS, method = "pearson")

rcorr(as.matrix(DesMx_combined_BUS),type = "pearson")
```

###Factors correlation

```
DesMx_com_F <- read_excel("DESMX_com_F.xlsx")

item_intercor(DesMx_com_F)

cor(DesMx_com_F, method = "pearson")

rcorr(as.matrix(DesMx_com_F),type = "pearson")
```

#Plots

```
PsyMx_Plot <- read_excel("PSYMX_Plot.xlsx")

PsyMx_Plot %>%
  ggplot(aes(x = UMUX_Lite, y = BUS_11)) +
  geom_point() +
  geom_smooth()

PsyMx_Plot %>%
  ggplot(aes(x = Familiarity, y = BUS_11)) +
  geom_point() +
  geom_smooth()

PsyMx_Plot %>%
  ggplot(aes(x = RSME, y = BUS_11)) +
  geom_point() +
  geom_smooth()
```