Developing and Testing the Russian Version of the BOT Usability

Wladimir Kukuruza

s2361264

w.kukuruza@student.utwente.nl

Department of Cognitive Psychology and Ergonomics

University of Twente

First supervisor: Dr. Borsci

Second supervisor: Jule Landwehr

July 1, 2022

Abstract

Chatbots are becoming increasingly important, especially in the customer service section. However, reliable and valid usability measurements tailored to chatbots are still a rarity. Usability can be measured by satisfaction. The "BOT Usability Scale" (BUS-11), developed by Borsci et al. (2021) attempts to do just that; measure user satisfaction towards chatbots. While this questionnaire is promising, there are only a few translated versions of it Borsci et al. (2021). Therefore, not all countries and its populations can use this tool to improve chatbots. To increase the number of translated versions of this questionnaire, the BUS-11 was translated into the Russian language and tested for its validity by performing a confirmatory factor analysis. To further investigate the validity, a correlation analysis with the translated version of the UMUX-LITE questionnaire was performed. Forty-three participants interacted with six chatbots and completed one task per chatbot. After interacting with each chatbot, they filled out the translated version of the BUS-11 (RUS-BUS-11) and the translated version of the UMUX-LITE (RUMUX-LITE). The PCA confirmed the five-factor model of the RUS-BUS-11. Furthermore, the correlation analysis with the RUMUX-LITE showed a high and significant correlation with the RUS-BUS-11, indicating high validity. Moreover, to ensure a correct translation of the RUMUX-LITE, the reliability of this scale was also investigated. The results showed that the translated version is as reliable as the original. Lastly, the influence of age on user satisfaction towards chatbots was assessed. A simple linear regression showed no significant relationship between age and user satisfaction towards chatbots. The results suggest that the RUS-BUS-11 is a reliable and valid tool to measure user satisfaction in the Russian language and the RUMUX-LITE is also a reliable measurement tool for measuring satisfaction towards chatbots. However further evaluation of the RUS-BUS-11 is recommended, due to the sample method and size and because of an accidentally mistranslated factor.

Keywords: Chatbots, Conversational Agents, BUS-11, Satisfaction, UMUX-LITE

Table of contents

1. Introduction

Chatbots are computer programs that understand and use natural language to communicate with users (Radziwill & Benton, 2017) by speech and text (McTear, 2017). The goal of the chatbot is to analyse the user´s input and give an appropriate answer that helps the user to answer his/her questions (Neumeister, 2020). Although the popularity of conversational agents has risen in the last two decades (Dale, 2016), chatbots have a much longer history. One of the first chatbots was developed by Joseph Weizenbaum, which he presented to the public in 1966: ELIZA. Weizenbaum´s aim was to create a chatbot that could act like a therapist and give appropriate reactions (Ireland, 2012). This chatbot was able to respond so accurately to the user's input that Weizenbaum's secretary believed ELIZA was a real person (Weizenbaum, 1976). Most chatbots during this time were a tool that helped to create artificial intelligence. One example on how chatbots were used for this is the Turing test. The aim of this test was for the participant to identify whether he/she was interacting with a human or a robot (Saygin, Cicekli, & Akman, 2000). In other words, when the participant cannot identify whether he/she speaks to a computer or a human, the computer wins. Although the test is now more than 70 years old, 2014 was the first time a chatbot passed the Turing test (Warwick & Shah, 2015). Around this time the usage and quality of chatbots had increased (McTear, 2017).

Recent developments in artificial intelligence improved the quality of conversational agents by refining natural language processing and machine learning (Gnewuch et al., 2017; Skjuve & Brandtzæg, 2019). In addition to the increased quality of chatbots, another reason for the growing prevalence of chatbots is the growing use of text-based communication tools such as SMS, email, etc. According to Dale (2017), 7.3 billion people own a mobile phone

that is capable of text-based communication. Even though speech-based chatbots are used as well (Gnewuch et al., 2017), conversational agents that use text as the base of communication are more commonly used (Araujo, 2018). Because of the prior mentioned arguments and the increased usage of the Internet worldwide (Boot et al., 2012), chatbots became increasingly useful in numerous fields from therapeutic settings to the customer service section (Araujo, 2018; Skjuve & Brandtzæg, 2019).

With the increasing number of internet users, the demand for customer service increases as well. Especially in this sector, the use of chatbots has increased significantly (Gnewuch et al., 2017). An advantage of conversational agents regarding the usage in the context of customer service is their time-saving and cost-effectiveness (Gnewuch et al., 2017). This is because chatbots simplify information retrieval, mainly for people unfamiliar with the internet (Gnewuch et al., 2017) and consequently help users find information more efficiently even on more complex websites (Jenkins et al., 2007). Therefore, an efficient and effective chatbot can reduce the amount of customer service calls and text replies from employees. Because of the advantages that chatbots can provide, it is expected to see an increased number of chatbots (Araujo, 2018).

Despite the recent improvements in the quality of chatbots, there is still a gap between what users expect from chatbots and what they are able to do. According to Gnewuch et al. (2017), despite the latest refinements of the chatbot quality, most chatbots still did not meet the customers' expectations. Kvale et al. (2021) discovered that dissatisfaction with the skills of the chatbots explains a large proportion of the perceived usability. A reason for chatbots not matching the expectations of users is that chatbots tend to reply in a way that is not context related. This can occur when for example, the user makes grammatical errors or uses everyday speech (Nuruzzaman & Hussain, 2018). In addition, users have a high expectation towards chatbots and therefore expect high-quality service (Kim et al., 2003). Therefore,

when users have an encounter with a chatbot that does not meet the expected criteria, people´s satisfaction with the chatbot decreases. Even though the quality of chatbots has increased in recent years and the demand for them is higher than ever, the user expectation is still not met.

To meet the expectations of the users, a reliable and valid way to measure the user experience towards chatbots will help designers to create better chatbots. User experience is defined by the international standard of human-centered design (ISO 9241-210), as a "person´s perceptions and responses resulting from the use and/or anticipated use of a product, system or service" (cited in Følstad & Brandtzæg, 2020, p.3). Two aspects have been shown to be a major influence in user experience; usability and usefulness (Følstad & Brandtzæg, 2020). Usefulness is important to investigate the value a tool has for the completion of a task and usability describes the effectiveness, efficiency and satisfaction of a task accomplishment (Tsakonas and Papatheodorou, 2008). Therefore, one way to measure the user experience is through investigating usability and usefulness. Tsakonas and Papatheodorou (2008) showed that satisfaction is related to both aspects and is therefore suitable to measure user experience (Følstad & Brandtzæg, 2020).

Consequently, satisfaction measurements help to detect possible flaws in chatbots. Short scales that explore the dimensions of usability and satisfaction already exist, like the UMUX (Bosley, 2013), the UMUX-LITE (Lewis et al., 2013) and the SUS (Brooke, 1996), that measure satisfaction through a list of questions. These questionnaires showed to be equally valuable measurement tools for measuring satisfaction because these scales were found to have a high correlation with each other (Borsci et al., 2015). Van den Bos and Borsci (2021) mentioned that these usability scales are not especially made for evaluating the perceived usability towards chatbots. Even though the scales showed a high reliability in general, chatbots differ from other technologies in their characteristics and are therefore not fit sufficient to evaluate user satisfaction towards chatbots (van den Bos & Borsci, 2021).

When using an overall usability measurement tool, researchers are less likely to pinpoint the exact reason why a chatbot is particularly usable and which aspects could be improved (Tariverdiyeva & Borsci, 2019). For these reasons, user satisfaction measurement tools designed for chatbots are needed to improve chatbots.

As mentioned above, current questionnaires do not take specific characteristics of chatbots into account. As a result, present scales like the SUS are unable to capture the full picture of chatbots (Følstad et al., 2018; van den Bos & Borsci, 2021). Consequently, Borsci et al. (2021) created a questionnaire that should measure user satisfaction with chatbots. Eventually, a questionnaire with 15 items with high reliability between .76 and .87 was developed. This *BOT Usability Scale* (BUS-15) consists of five factors and showed a strong correlation with the already validated UMUX-LITE questionnaire. A five-point Likert scale was used to assess the participants' agreement with each item ranging from 1 (‘Strongly Disagree’) to 5 (‘Strongly Agree’). However, after performing confirmatory factor analysis, Borsci et al. (2021) shortened the questionnaire by identifying the most significant items. As a final version, an 11-item questionnaire composed of five factors with the name *BOT Usability Scale* (BUS-11) with overall reliability of .9 (Table 1) was introduced. This new scale showed also a high correlation with the UMUX-LITE where the psychometric properties were already confirmed. This newly created scale also used a five-point Likert scale ranging from 1 (‘Strongly Disagree’) to 5 (‘Strongly Agree’). Even though the BUS-11 seems to be a reliable measure for assessing user satisfaction with chatbots, this questionnaire exists in a limited number of languages (Borsci et al., 2021). As a result, not all chatbots that have a different language can be reviewed.

Table 1

11-item BOT Usability Scale (BUS-11)

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

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

1.1 Aim of this study

To investigate whether the RUS-BUS-11 can be used as a valid and reliable tool to measure user satisfaction towards chatbots for Russian speaking websites, this study aims to perform a confirmatory factor analysis to investigate the psychometric properties of the RUS-BUS-11. *Research question Q1. Are the factorial structure and the psychometric (i.e., factorial structure and reliability) properties of RUS-BUS-11 in line with previous studies (Borsci et al., 2021)?*

As shown in the previous study, the BUS-11 and the UMUX-LITE showed a strong correlation (Borsci et al., 2021). In order to investigate the convergent validity of the RUS-BUS-11, the UMUX-LITE will be translated into Russian. In the further, the translated version of the UMUX-LITE will be called RUMUX-LITE. *Research question Q2. Does the RUS-BUS-11 correlate with the RUMUX-LITE?*

In order to be able to compare the translated version of the BUS-11 with the UMUX-LITE, it is advantageous to translate this questionnaire as well. To ensure a correct translation of the UMUX-LITE, the reliability of this scale has to be investigated, to find out if this questionnaire measures what it is supposed to measure (Mcleod, 1970). Previous studies have shown that the UMUX-LITE had an acceptable reliability ($\alpha = 0.7$) (Borsci et al., 2021). Therefore this study will investigate the reliability of the translated version of the UMUX-LITE. *Research question Q3. Does the RUMUX-LITE have comparable reliability to the UMUX-LITE*?

In the exploratory study that originated the BUS scale, Borsci et al. (2021) suggested that there is a need for a more diverse range of people. More diversity of ability, age and gender was pointed out. Moreover, results of previous analyses (Borsci et al 2021) suggested that there is an effect of age on satisfaction. Therefore, this study will also investigate whether

there is an effect of age on satisfaction toward chatbots using the RUS-BUS-11. *Research question Q4. Does age affect satisfaction with chatbots rated by the RUS-BUS-11?*

2. Methods

2.1 Participants

Ethics approval of the Ethics Committee of the University of Twente was obtained before the recruitment of participants. The participants were recruited through the snowballing sampling method and the convenience sampling method. Participants were found by recruiting in the private and social networks (e.g., Reddit) of researchers aiming for Russian speakers. Reddit was chosen because it offers users to communicate in various forums called “subreddits”. One of these groups aims to provide a platform where users who learn Russian can interact with users who are Russian speakers. Since this subreddit, called "r/Russian", is about the Russian language in general, Reddit was used to post an announcement in this group. In addition, the subreddit “r/Participants” was used to find candidates for the survey. For the same reason, the social media platform Facebook was used. A notice was posted in the forums "Survey Exchange/ Survey Group/ Survey Participants" and "SurveyCircle/Survey Panel". In the SONA system of the University of Twente was also used to get more participants into the study. Students at the University of Twente were compensated with 0.25 SONA points. Before participating, participants were obligated to read the information sheet and to agree with the consent form (see Appendix A). Sixty-five volunteers participated. Three of them were excluded because there was no variance in their answers. This indicates that these three did not complete the task but clicked their way through to the end. Twenty participants did not give their consent and therefore were also excluded. In the end, 42 completed the survey (*M*age=38.59 ;*SD*age=15.5) with a range between 18 and 70 years. Eighteen of them were male and 24 female. The majority of participants were Ukrainian (42.85%), while 19.04% were Russian, 14.29% German, three were Belarussian, two Kazakhstanis, three Israelis, one

Hungarian and one from Estonia. Furthermore, 73.33% stated that Russian was their native language. 17.78% said that Ukrainian was their native language. Two participants´ mother language was Belarusian and two stated it was Kazakh. Lastly, participants were asked about their proficiency in the Russian language. 71.11% stated that their Russian language skills were at the C2 level. 20% had a C1 level, and 8.88% had a B2 level.

2.2 Materials

The software Qualtrics (n.d.) was used to gather data on the online assessment of chatbots. Informed consent was displayed as a prerequisite to start with the online assessment. A short demographics form was provided that included age, gender, nationality and their mother tongue (Appendix B). To check the language proficiency in Russian, a 6-point Likert scale was used ranging from “Beginner” as the lowest level to “Proficient/Native Speaker” as the highest level (Appendix B). Because of the rarity of chatbots in the Russian-speaking world and the current trend of the life-chat function with the customer service, five chatbots could be found in total. Therefore, five Russian websites with chatbots were included in the study (Appendix C). Qualtrics (n.d.) was also used for the tasks to provide participants with a link to the websites with the associated chatbots. Per each chatbot, participants were asked to perform one task (Appendix C). Every participant interacted with these five chatbots. The chatbots were presented in random order. To make sure that every task had approximately the same difficulty and are comparable to one another, tasks were selected that could be completed by asking between 2 to 3 questions to the chatbots. Therefore participants could complete the task and get the chatbot to reveal the necessary information by asking two to three questions for each task.

A native speaker with a B1 level in English translated the BUS-11 and the UMUX-LITE into the Russian language (see Appendix D). In order to reveal potential misunderstandings and to ensure the accuracy of the translation, the initial translation was

back-translated into English by another Russian native speaker with the same English proficiency (see Appendix E). After comparing the re-translated version and the original version, the translation of the scales was considered accurate. The RUS-BUS-11, similarly to the original version is composed of a Five-point Likert scale that ranged from “strongly disagree” to “strongly agree”. This Questionnaire consisted of 11 questions, split up into five factors. Factor 1 “Perceived accessibility to chatbot functions” is composed of items one and two. In Factor 2, Items three, four and five were about the perceived quality of chatbot functions. Item six to nine asked about the perceived quality of conversation and information provided. Ten was about perceived privacy and security and item 11 measured the time response.

Regarding the RUMUX-LITE, this scale was composed of a 7-point Likert Scale from “strongly disagree” to “strongly agree” with a minimum and maximum score for every participant for each chatbot from 2 to 14 was used. The UMUX-LITE consists of two items; “This system’s capabilities meet my requirements” and “This system is easy to use”. This questionnaire undertook the same translation procedure as the RUS-BUS-11. To make sure that participants understood the tasks, a pilot trial was performed. Two participants were asked to perform the tasks for each chatbot and report whether they understood the instructions. After performing the pilot study, no errors were found. Therefore, the study was considered functional and was published after the review by the Ethics Committee of the University of Twente.

2.3 Procedure

Respondents who were interested in participating in the survey got an E-mail invitation to the survey or were provided with an anonymous link. After following the link, participants were instructed to read and accept the consent form and to provide information about their demographics. Participants filled out the six-point Likert scale about their proficiency in the Russian language. Participants were presented first with the name of the website and a link was provided where the participant was instructed to go into the website. Beneath, a scenario was presented to better understand the task at hand (Appendix C). Arriving at the website, the participants first had to find the chatbot before they could interact with it. If the respondents felt they had completed the task or if they perceived that the task was unachievable, they were asked to return to the Qualtrics website to complete the RUS-BUS-11 and the UMUX-LITE. After the completion of the survey, participants had to interact with four more different chatbots following the same scheme mentioned above. At the end of the assessment, participants were thanked and were provided with the contact information of the researcher to be able to ask questions and provide feedback.

2.4 Data analysis

The gathered data from the survey in Qualtrics (n.d.) was exported to Microsoft Excel 365. The data of participants who did not complete the survey were deleted. In addition, participants who had a lower Russian language skill level than B2 were taken out.

During the data analysis, it was noted that item 10 (originally: “I believe the chatbot informs me of any possible privacy issues”) was translated in a way that could be understood as “I believe that the chatbot saves my private data”. This imprecise interpretation could lead participants in understanding the statement as the opposite of its actual meaning. The translated version of item 10 cannot represent the original item 10, and therefore could influence the results. The wrongly translated item was still used in this research to investigate

whether the questionnaire is still reliable and valid. The correct translation of item 10 (“Я считаю, что чатбот сообщает мне о возможных проблемах с конфиденциальностью”) should be used to further investigate the psychometric properties of the Russian version of the scale. To partially compensate for this translation error the score of item 10 was reversed. For example, when a participant scored five points in the given statement, “I believe that the chatbot saves my private data”, it was turned into one point.

The data obtained by Qualtrics (n.d.) were renamed and imported to the statistical program R (R Core Team, 2021). In addition, to perform the confirmatory factor analysis, the ‘lavaan’ package by Rosseel et al. (2021) was installed as well. To represent the factor model graphically, the R package “semPlot” (Epskamp et al., 2022) was used.

To investigate the psychometric properties of the RUS-BUS-11 and the factorial structure of the scale in line with the previous study (Borsci et al., 2021), a confirmatory factor analysis (CFA) was performed whereby the parameters of the model were specified to match the original 5 factors of the BUS-11 (Table 1). The Shapiro-Wilk test was used to test the assumption of normality. The items with $p > .05$ were considered to have a normal distribution (Hanusz et al., 2014). For non-normal distributed items, the robust maximum likelihood was used to run the CFA (Li, 2016). To evaluate the model fit, the following indexes were considered: Chi-square with $\chi 2 > 0.01$, the comparative fit index (CFI) with CFI≥ 0.9, the Tucker Lewis Index (TLI) with TLI ≥ 0.95, the root mean square error of approximation (RMSEA) with RMSEA ≥ 0.05 and ≤ 0.08 for acceptable fit and the standardized root mean square residual (SRMR) with SRMR≤ 0.05 for good fit (Barney et al., 2021; Harerimana & Mtshali, 2020). Standardized factor loadings and the variance of the items were calculated and examined to investigate the correlation between the items and their factors and the proportion of the variance explained by the model. Factor loadings of >.6 are seen as acceptable (Peterson, 2000).

The reliability of the RUS-BUS-11, as well as the RUMUX-LITE, was estimated using the Cronbach's Alpha by the R package "ltm" (Rizopoulos, 2007). An Alpha value above .70 is seen as acceptable and a value of .80 or greater is seen as preferred (Cortina, 1993).

The average scores for the RUMUX-LITE and the RUS-BUS-11 were calculated to perform the Kendall´s Tau test to establish the convergent validity of the RUS-BUS-11 with the RUMUX-LITE.

Finally, to explore whether age affects the overall satisfaction of chatbots, a regression analysis was performed with age as the independent variable, and the overall score of the RUS-BUS-11 as the dependent variable.

3. Results

3.1. Psychometrics properties of the Russian version of the BUS-11

To answer the first research question Q1: *Are the factorial structure and the psychometric (i.e., factorial structure and reliability) properties of RUS-BUS-11 in line with previous studies (Borsci et al., 2021)?;* a confirmatory factor analysis was performed. First, the assumptions of normality of each BUS item using the Shapiro-Wilk test were assessed. All items were considered as nonnormal distributed ($p < .001$). Based on this outcome, the maximum likelihood (MLR) method was used. Overall, the five-factor model of the BUS11 showed a good fit as reported in Table 2.

All indicators showed solid factors loading (Table 3) except for Item 1 which presents a problem associated with the variance. The other items had a standardized factor loading of 0.6 or higher and therefore contributed heavily to their factors, while the monodimensional items (10 and 11) had a standardized factor loading of 1 as expected.

Table 2

Fit indices from the confirmatory factor analysis of the RUS-BUS-11 model

	RUS-BUS-11	
$\chi 2$ ($p > .05$)	78.101	$p < .001$
RMSEA $\geq .05$ and $\leq .08$	.085	
SRMR ($\leq .05$)	.038	
CFI ($\geq .9$)	.972	
TLI ($\geq .95$)	.957	

Note. The fit indicators are Chi-square goodness of fit statistics ($\chi 2$), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), and Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI).

Table 3

Factor Loadings of the RUS-BUS-11

Factor	Item	Std. Factor Loading	Std. Variance	Variance – Lower Bound	Variance – Upper Bound
1: Perceived accessibility to chatbot functions	1	1.04	-.076	-.279	.069
	2	.77	.409	.359	.977
2: Perceived quality of chatbot functions	3	.85	.277	.274	.585
	4	.85	.282	.342	.663
	5	.89	.204	.211	.451
3: Perceived quality of conversation and information provided	6	.89	.212	.291	.547
	7	.95	.094	.118	.280
	8	.91	.175	.213	.464
	9	.87	.247	.295	.589
4: Perceived privacy and security	10	1	0	0	0
5: Time response	11	1	0	0	0

Note. This Table demonstrates the standardized factor loadings, the standardized variance, and the confidence intervals of unstandardized variance of the RUS-BUS-items

3.2 Reliability of the RUS-BUS-11

The internal consistency of the RUS-BUS-11 was assessed by performing a reliability analysis. The Cronbach´s alpha was calculated and the results showed that the RUS-BUS-11 has a good internal consistency ($\alpha = 0.908$).

3.3 Correlation between RUMUX-LITE and the BUS-11

To answer the second research question Q2: *Does the RUS-BUS-11 correlates with the RUMUX-LITE;* first, a normality test was performed on the overall scores of the RUS-BUS-11 ($W= 0.907$, $p < 0.001$) as well as of the RUMUX-LITE ($W= 0.887$, $p < 0.001$). This suggested that the two scores were nonnormally distributed. Kendall's Tau test indicated that the results of the two scales are strongly correlated ($\tau b=0.585$, $p<0.001$). In addition, the relationship between the RUMUX-LITE and each factor of the RUS-BUS-11 was

investigated (Table 4). RUMUX-LITE had a significant positive correlation with factors F1, F2, F3 and F5. Factor F4 had a significant negative correlation with the RUMUX-LITE (Table 4).

Table 4

Correlations between the RUMUX-LITE and the RUS-BUS-11

	RUMUX-LITE
RUS-BUS	.585***
(f1) Perceived accessibility to chatbot functions	.455***
(f2) Perceived quality of chatbot functions	.590***
(f3) Perceived quality of conversation and information provided	.511***
(f4) Perceived privacy and security	-.228***
(f5) Time response	0.373***

Note. Correlations were measured with Kendall's rank correlation method. ****p< 0.001*

3.4 Reliability of the RUMUX-LITE

To answer the third research question Q3: *Does the RUMUX-LITE has the same reliability as the UMUX-LITE*?; Cronbach´s alpha was calculated. The analysis showed a good internal consistency of the RUMUX-LITE ($\alpha = 0.874$) indicating a good model fit.

3.5 Effect of age on the satisfaction of chatbots

To explore the effect of age on satisfaction (Research question Q4) linear regression analysis was performed. The results indicated a non-significant correlation ($F(208)= 0.315$, $p= 0.575$).

4. Discussion

This study aimed to test the psychometric properties of the Russian-translated version (RUS-BUS-11) of the BUS-11 chatbot satisfaction scale recently developed by Borsci et al. (2021). In addition, to validate the scale, a correlation analysis was performed between the RUS-BUS-11 and the translated version of the UMUX-LITE (RUMUX-LITE). Next, the internal consistency of the RUMUX-LITE was investigated. Furthermore, a regression analysis was done to investigate whether age affects user satisfaction with chatbots.

To answer the first research question Q1: *Are the factorial structure and the psychometric (i.e., factorial structure and reliability) properties of RUS-BUS-11 in line with previous studies (Borsci et al., 2021)?;* the reliability and the validity of the scale were examined. The validity was investigated using the confirmatory factor analysis. The results could confirm the five factor model that was developed by Borsci et al. (2021). These factors are perceived accessibility to chatbot functions, perceived quality of chatbot functions, perceived quality of conversation and information provided, perceived privacy and security, and time response (Table 1).

Overall, the results indicated a good model fit, nevertheless, the value of the root mean square error of approximation (RMSEA) is above the threshold. It has been shown in the past that the RMSEA has serious problems with simpler models with few degrees of freedom. This is especially true for simple path models and simple CFAs, which more often have relatively few degrees of freedom. Here, the RMSEA can falsely indicate a poor fit, even when in fact the model fits the data well (Kenny et al., 2015). Another explanation for a high RMSEA is that the wrong translation of item 10 could have affected the quality of the model. Even though the question has been translated so that it has a different meaning than the original one (originally: “I believe the chatbot informs me of any possible privacy issues” and the current item could be understood as “I believe that the chatbot saves my private data”), The results

have nevertheless had no abnormalities for item 10. Accordingly, the results showed that item 10 measured the attributed factor 4: perceived privacy and security.

Overall the present version of RUS-BUS-11 is fitting the expected model quite well, however, future studies should involve a larger number of participants to further explore the issues with the RMSEA we reported. Furthermore, item 1 showed a standardized factor loading that is exceeding the maximum value of one (see Table 4). This is an indicator of the so called “Heywood case”. According to the American Psychological Association (n.d.), a sample that is too small or data that is nonnormal distributed can lead to factor loadings that have impossible or very rare values. Due to this, it is recommended to select a sample large enough to adequately estimate the parameters to avoid standardized factor loadings that exceed the maximum values of one.

Despite these issues, the RUS-BUS-11 reliability was good and we can report that, overall, the present Russian version of the scale is satisfactory in line with the original version proposed by Borsci et al. (2021). Furthermore, by performing a correlation analysis, our results suggested that the RUS-BUS-11 and the RUMUX-LITE are strongly and positively correlated and therefore being in line with the second research question; “*Does the RUS-BUS-11 correlate with the RUMUX-LITE?”.* This relationship between the two scales and are therefore in accordance with the previous results of Borsci et al. (2021).

These results suggest that the Russian version of the scale supports the psychometric properties of the original scale. A valid and reliable translation of a scale can help to gather data and compare it between cultures and therefore can help to illustrate cultural and linguistic differences between populations (Yu et al., 2004). Differences in responses can be seen when comparing answers to the same questionnaires in different languages. This indicates that correctly translated questionnaires help to accurately capture the answers of participants independent without the influence of the respondents' native language and cultural

background (Harzing, 2005). Based on this, it was confirmed that the RUMUX-LITE questionnaire is an adequate measurement tool, which can help to test Russian speaking chatbots for their usability.

Previous studies suggested an effect of age on user satisfaction with chatbots (Borsci et al., 2021). As a consequence, this study aimed to answer the research question Q4: *Does age affects satisfaction with chatbots rated by the RUS-BUS-11.* The results showed a non-significant effect on the relationship between age and the total score of the RUS-BUS-11. That being the case, it can be concluded that in this study, age did not predict the outcome of user satisfaction with chatbots. Therefore, the results are not in line with the previous indications (Borsci et al., 2021). Moreover, Borsci et al. (2021) stated that: "…a more diverse range of people (age, gender and ability) are needed to use the system in future iterations; in this study mainly young participants with age below 35 years old were involved in focus groups and in the pilot of the scale" (p.15). Because this study included a wider range of ages and a higher mean age, the results of this study can support the assumption that the BUS-11 is a reliable tool.

4.1 Limitations and future research

The first limitation of the current study was the amount of chatbots in the Russian speaking world. Only five chatbots were found on Russian-speaking websites. In the previous study by Borsci et al. (2021), it was stated that a larger number of chatbots are needed to ensure the reliability of the construct. Consequently, due to the limited number of chatbots used in this study, it is recommended to add more chatbots in the next study if possible.

Another concerning issue was the sample and the sample size that was used. The snowball sampling method was used instead of the preferred probability sampling method because of the time constraints and the need to recruit Russian speakers. This could lead to biases in the sampling data because the majority of participants were in the social network of

the researcher, therefore the findings can only cautiously be inferred from the population because the sample probably does not reflect all the characteristics of the various social groups present in the population (Etikan et al., 2018). The snowballing sampling method is problematic because friends and family members tend to share the same traits and characteristics. Therefore, the sample does not represent the whole diversity of the population (Biernacki & Waldorf, 1981). In addition, because the sample size has a significant effect on the confirmatory factor analysis, the sample size could have influenced the results. In addition, the standardized factor loading of item 1 was above the maximum value of one. This could be attributed to sampling fluctuations (Kolenikov & Bollen, 2012). Therefore it is recommended for future research to use a probability sampling method and use a greater sample size in order to prevent potential sample biases and effects on the analysis.

Another potential limitation was the current Covid-19 pandemic. Participants did the study online. Therefore it was not possible to ensure that the participants did the survey in the same standardized and controlled environment. It was not possible to detect the influence of potential disturbances that might have occurred during the study for example noise disturbances. Furthermore, participants could have used different tools e.g., laptops, mobile phones or tablets during the survey. This could have altered the way the website looks and the difficulty to find and interact with the chatbot. In addition, the internet connection could have influenced the results, too. Participants with a slower internet connection might answer the questionnaire differently than participants with a faster internet connection. Therefore, it is recommended for future research to provide a standardized and controlled environment for every participant to reduce the likelihood of such influences.

It is recommended to pay special attention to the RMSEA score to see if the implementation of the correctly translated item affects the score. The translation error that occurred for item 10 (see Appendix D and E) could have influenced the understanding of the

item and therefore influenced the overall score and therefore the results. Despite this translation error, the current item 10, “I believe that the chatbot saves my private data” seemed to measure factor 4, therefore it is recommended for future research to investigate whether correctly translated version of item 10 that was mentioned earlier will fit better than the wrongly translated version of item 10.

In addition, according to Country Comparison (2021), eastern European culture differs from western European in some points, for example in “Uncertainty avoidance”. This could have influenced the way people scored for example on factor 4 (“perceived privacy and security”). Further studies should take the corrected version of the RUS-BUS-11 and it is recommended to investigate whether specific cultural characteristics influence the results.

5. Conclusion

Because of the need for more questionnaires that measure satisfaction towards chatbots, the current study investigated the psychometric properties of the translated version of the BUS-11, the RUS-BUS-11. This study has proven that the translated version of BUS-11 is valid and reliable. Furthermore, the translated version of the UMUX-LITE was shown to be reliable and therefore both questionnaires can be used in the future to improve the quality of chatbots. Lastly, this study investigated the relationship between age and user satisfaction with chatbots and found no significant influence of age on user satisfaction with chatbots. Moreover a bigger sample size and a probability sampling method is recommended for future studies in addition to a controlled environment for every participant.

References

American Psychological Association. (n.d.). *Apa Dictionary of Psychology*. American Psychological Association. https://dictionary.apa.org/heywood-case

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

Barney, J. L., Barrett, T. S., Lensegrav-Benson, T., Quakenbush, B., & Twohig, M. P. (2021). Confirmatory factor analysis and measurement invariance of the cognitive fusion questionnaire-body image in a clinical eating disorder sample. *Body Image*, *38*, 262–269. https://doi.org/10.1016/j.bodyim.2021.04.012

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research, 10*(2), 141–163. https://doi.org/10.1177/004912418101000205

Boot, W. R., Nichols, T. A., Rogers, W. A., & Fisk, A. D. (2012). Design for aging. *Handbook of Human Factors and Ergonomics*, 1442–1471. https://doi.org/10.1002/9781118131350.ch52

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), 484495.

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

Bosley, J. J. (2013). Creating a short usability metric for user experience (UMUX) scale. *Interacting with Computers, 25*(4), 317-319

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Country comparison. Hofstede Insights. (2021). Retrieved from https://www.hofstede-insights.com/country-comparison/

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

Epskamp, S., Stuber, S., Nak, J., Veenman, M., & Jorgensen, T. D. (2019). semPlot: Path diagrams and visual analysis of various SEM packages' output. Retrieved from https://cran.r-project.org/web//packages/semPlot/semPlot.pdf

Etikan, I., Alkassim, R., & Abubakar, S. (2015). Comparision of Snowball Sampling and Sequential Sampling Technique. *Biometrics & Biostatistics International Journal*, *3*(1), 00055. DOI: 10.15406/bbij.2015.03.00055

Følstad, A., & Brandtzæg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience 5*(3), 1-14. https://doi.org/10.1007/s41233-020-00033-2

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In *International Conference on Internet Science*, 194-208. Springer, Cham.

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems (ICIS)*, 1-13. Retrieved from https://www.researchgate.net/profile/UlrichGnewuch/publication/320015931_Towards_Designing_Cooperative_and_Social_Conversational_Agents_for_Customer_Service/links/59c8d1220f7e9bd2c01a38a5/Towards-Designing-Cooperative-and-Social-Conversational-Agents-for-Customer-Service.pdf

Hanusz, Z., Tarasinska, J., & Zielinski, W. (2016). Shapiro-Wilk test with known mean. *REVSTAT 14*(1), 89–100. Retrieved from https://www.ine.pt/revstat/autores/pdf/rs160105.pdf

Barney, J. L., Barrett, T. S., Lensegrav-Benson, T., Quakenbush, B., & Twohig, M. P. (2021)

Harerimana, A., & Mtshali, N. G. (2020). Using Exploratory and Confirmatory Factor Analysis to understand the role of technology in nursing education. *Nurse Education Today*, *92*, 104490

Harzing, A.-W. (2005). Does the Use of English-language Questionnaires in Cross-national Research Obscure National Differences? *International Journal of Cross Cultural Management, 5(*2), 213–224. https://doi.org/10.1177/1470595805054494

Ireland, C. (2012). *Alan Turing at 100*. Retrieved from https://news.harvard.edu/gazette/story/2012/09/alan-turing-at-100/

Jenkins, M.-C., Churchill, R., Cox, S., & Smith, D. (2007). Analysis of user interaction with service oriented chatbot systems. *Proceedings International Conference of Human-Computer Interaction*, 76–83. https://doi.org/10.1007/978-3-540-73110-8

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

Kim, B., Park, K., & Kim, J. (2003). Satisfying different customer groups for IS outsourcing: A korean IS company's experience. *Asia Pacific Journal of Marketing and Logistics,15*(3), 48–69. doi:10.1108/13555850310765006

Kolenikov, S., & Bollen, K.A. (2012). Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?. *Sociological Methods & Research 41*(1) 124–167. DOI: 10.1177/0049124112442138

Kvale, K., Freddi, E., Hodnebrog, S., Sell, O. A., & Følstad, A. (2021). Understanding the user Experience of Customer Service Chatbots: What Can We Learn from Customer Satisfaction Surveys?. In: *CONVERSATIONS 2020, LNCS 12604*, 205–218. https://doi.org/10.1007/978-3-030-68288-0_14

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of CHI 2013* (pp. 2099–2102). Paris, France: ACM. doi:10.1145/2470654.2481287

Li, C. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum Likelihood and diagonally weighted least squares. *Behav Res, 48*, 936–949. DOI 10.3758/s13428-015-0619-7

Mcleod, S. (1970). *What is reliability?* What is Reliability? | Simply Psychology. Retrieved from https://www.simplypsychology.org/reliability.html

McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? *Lecture Notes in Computer Science*, 38–49. https://doi.org/10.1007/978-3-319-69365-1_3

Neumeister, S. (2020). *Testing of a usability assessment tool for chatbots: investigating the effect of believing that a chatbot might be a human* (Bachelor's thesis, University of Twente).

Nuruzzaman, M., & Hussain, O. K. (2018). A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. In *IEEE 15th International Conference on e-Business Engineering (ICEBE)*, 54-61. DOI 10.1109/ICEBE.2018.00019

Peterson, R. A. (2000). A Meta-Analysis of Variance Accounted for and Factor Loadings in Exploratory Factor Analysis. *Marketing Letters 11*(3), 261-275. Retrieved from https://link.springer.com/content/pdf/10.1023/A:1008191211004.pdf

Qualtrics (n.d.). *Qualtrics XM* [Computer Software]. Retrieved from https://www.qualtrics.com

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Rizopoulos D (2006). "ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses." *Journal of Statistical Software*, **17**(5), 1–25. Retrieved from https://doi.org/10.18637/jss.v017.i05.

Rizopoulos, D. (2007). ltm: An R package for latent variable modeling and item response analysis. *Journal of statistical software*, *17*, 1-25.

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

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and machines*, *10*(4), 463-518. doi:10.1023/A:1011288000451

Skjuve, M. B., & Brandtzaeg, P. B. (2019). Measuring user experience in chatbots: An approach to interpersonal communication competence. In *International Conference on Internet Science*, 113–120. https://doi.org/10.1007/978-3-030-17705-8_10

Tariverdiyeva, G., & Borsci, S. (2019). Chatbots' perceived usability in information retrieval

tasks: An exploratory analysis. [Master Thesis]. University of Twente, Enschede, The Netherlands. http://essay.utwente.nl/77182/

Tsakonas, G., & Papatheodorou, C. (2008). Exploring usefulness and usability in the evaluation of open access digital libraries. *Information processing & management*, *44*(3), 1234-1250

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

Warwick, K., & Shah, H. (2015). Can machines think? A report on Turing test experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence*, *28(6), 989-1007.* doi:10.1080/0952813x.2015.1055826

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* W. H. Freeman & Co.

Yu, D. S. F., Lee, D. T. F., & Woo, J. (2004). Issues and Challenges of Instrument Translation. Western Journal of Nursing Research, 26(3), 307–320. https://doi.org/10.1177/0193945903260554

Appendices

Appendix A: Consent Form

Уважаемый участник. Мы приглашаем вас принять участие в научном исследовании. Участие в программе является полностью добровольным. Если вы согласитесь участвовать сейчас, вы всегда сможете отказаться и выйти из проекта. Негативных последствий не будет, что бы вы ни решили.

Цель исследования

Цель данного исследования - оценить опросник, предназначенный для измерения удовлетворенности пользователей чат-ботами.customer service. Чатбот - это программа, с которой вы можете общаться посредством текста, она дает ответы на ваши сообщения.

Содержание исследования

Вы получите несколько заданий, будете взаимодействовать с несколькими чатботами и после взаимодействия с каждым чатботом вам нужно будет пройти опрос для оценки чатбота. Исследование займет около получаса, и ваше участие в нем не связано с риском.

Сбор данных

В конце мы хотим использовать эти данные, чтобы узнать, какие вопросы действительно важны и помогут в оценке чатбота. Кроме того, перед началом опроса мы зададим несколько вопросов о вашем возрасте, поле, национальности и владении русским языком. Только руководители данного исследования смогут видеть эти данные. Возможно, данные будут опубликованы, но те из них, по которым вас можно будет идентифицировать, будут удалены. Информация будет храниться в защищенном хранилище данных университета, доступ к которому будет иметь только мой научный руководитель

Контакты

Если у вас возникнут вопросы после окончания этой сессии, вы можете написать мне по электронной почте: w.kukuruza@student.utwente.nl, а с моим научным руководителем можно связаться по адресу s.borsci@utwente.nl. С вопросами о ваших правах вы можете

обратиться по адресу ethicscommittee-bms@utwente.nl . Данное исследование одобрено этическим комитетом факультета поведенческих, управленческих и социальных наук (BMS) Университета Твенте.

Translation

Dear Participant. We invite you to participate in a scientific study. Participation in the program is completely voluntary. If you agree to participate now, you can always refuse and withdraw from the project. There will be no negative consequences, no matter what you decide.

Purpose of the Study

The purpose of this study is to evaluate a questionnaire designed to measure user satisfaction with chatbots. A chatbot is a program with which you can communicate via text, it replies to your messages.

The content of the survey

You will get several tasks, you will interact with several chatbots and after interacting with each chatbot you will have to take a survey to evaluate the chatbot. The survey will take about half an hour and there is no risk to your participation.

Data Collection.

At the end, we want to use this data to find out which questions are really important and will help in evaluating the chatbot. We will also ask some questions about your age, gender, nationality, and proficiency in Russian before we start the survey. Only the heads of this survey will be able to see this data. It is possible that the data will be published, but those by which you can be identified will be deleted. The information will be stored in a secure vault of the university, to which only my supervisor will have access

Contact

If you have any questions after this session, you can email me at w.kukuruza@student.utwente.nl, and my supervisor can be reached at s.borsci@utwente.nl. Questions about your rights can be directed to ethicscommittee-bms@utwente.nl . This

research is approved by the Ethics Committee of the Faculty of Behavioral, Management and Social Sciences (BMS) of the University of Twente.

Appendix B: Demographics Form

Какой пол вы имели при Рождении

- Мужской
- Женский
- Другое

Сколько вам лет? Пожалуйста, введите только цифры

Какая у вас национальность?

- украинец(ка)
- русский(ая)
- казах(шка)
- белорусс(ка)
- немец(ка)
- Другое

Какой ваш родной язык?

- Украинский
- Казахский
- Белорусский
- Русский
- Другое

Я считаю, что мой уровень русского языка...

- A1- Уровень Элементарного Общения
- A2- Предпороговый (базовый) Уровень
- B1- Пороговый Уровень
- B2- Постпороговый Уровень
- C1- Уровень Компетентного Владения
- C2- Уровень Носителя Языка

Translation

What gender did you have at birth

- Male
- Female
- Other

How old are you? Please enter only digits

What is your nationality?

- Ukrainian
- Russian
- Kazakh
- Belarusian
- German
- Other

What is your native language?

- Ukrainian
- Kazakh
- Byelorussian
- Russian
- Other

I believe that my Russian language level is...

- A1- Elementary level
- A2- Prerequisite (Basic) Level.
- B1- Threshold Level
- B2- Post Threshold Level
- C1- Competent Proficiency Level
- C2- Language Proficiency Level

Appendix C: Five Russian Websites with their chatbots

Chatbots and Tasks	Link
WTB Bank	https://www.vtb.ru/
Task: You live in a small apartment and your family going to have triplets. There is not enough room in the apartment for the whole family and you decided to buy a house. Question to use the bot: Find out if it's possible to buy a house with state support	
Aimylogic	https://aimylogic.com/ru
You are a private entrepreneur and have created your own online store. In order to simplify and reduce the time spent on correspondence with customers you have decided to install a chatbot on the site. Question for using the bot: Find out how to program a chatbot	
Eldorado	https://www.eldorado.ru/
You bought clothes in an online store. Unfortunately, it did not fit and you decided to return the purchased goods. Question to use the bot: Find out how many days the buyer has to return the product	
DNS-Shop	https://www.dns-shop.ru/
You are the manager of a company and want to buy goods from one of the online stores. Question for the chat bot: Find out what the offered delivery conditions are	
Belarusbank	https://belarusbank.by/ru/fizicheskim_licam/31886/internet_banking
You have decided to no longer rent housing for your family and buy an apartment. To do this, you need to take out a loan from a bank. Question to use the bot: Find out under what conditions can you get a loan to buy a home	

Note. This Table demonstrates the websites with chatbots and their assigned tasks that were used during this research. These tasks were translated. The original tasks can be found in Appendix B.

Appendix D: The RUS-BUS-11 and the RUMUX-LITE

Translated version of the 11-item BOT Usability Scale (BUS-11), the RUS-BUS-11

Фактор	**Параметры фактора**
1: Легкодоступность нахождения чатбота	1: Функция чатбота была легко распознаваема. 2: Чатбот было легко найти.
2: Интуитивное восприятие качества функциональности чатбота	3: Общение с чатботом было легко воспринимаемым 4: Чатбот был способен учитывать контекст 5: Ответы чатбота были понятны
3: Восприятие качества диалога и предоставляемой информации	6: Я нахожу, что чатбот понимает мои пожелания и помогает мне достичь цели. 7: Чатбот предоставляет мне необходимый объем информации. 8: Чатбот дает мне только ту информацию, которая мне необходима. 9: Я считаю, что ответы чатбота были точными.
4: Восприятие обеспечения конфиденциальности и защиты информации	10: Я уверен, что чат-бот сохраняет мои конфиденциальные данные
5: Время получения ответа	11: Время ожидания ответа от чат-бота было коротким.

Note. Item 10 was wrongly translated and should be translated as following: “Я считаю, что чатбот сообщает мне о возможных проблемах с конфиденциальностью”.

Translated version of the UMUX-LITE, the RUMUX-LITE

Возможности этой системы соответствуют моим требованиям
Эта система проста в использовании

Appendix E: Back Translation of the RUS-BUS-11 into English

Factor	Factor parameters
1: Ease of access a chatbot finding	1: The chatbot function was easily recognizable 2: The chatbot was easy to find
2: The intuitive perception of the quality of chatbot functionality	3: The communication with the chatbot was easy to understand 4: The chatbot was able to take into account the context 5: The chatbot responses were clear
3: The perception of the dialog quality and the provided information	6: I consider that a chatbot understands my wishes and helps me to achieve my goal 7: The chatbot provides me with the necessary amount of the information 8: The chatbot gives me only the information that I need 9: I believe the chatbot responses were accurate

4: The perception of privacy and information security provision	10: I believe that the chatbot saves my private data***
5: Response time	11: The response time from the chatbot was short

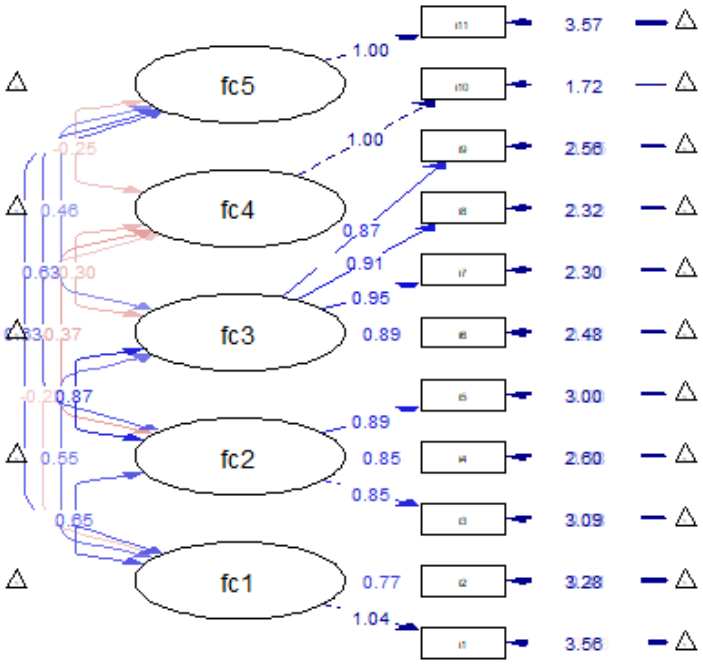
Note. Item 10 was wrongly translated and should be translated as following: “I believe that the chatbot saves my private data”.

*** During the backtranslation I made something wrong…but maybe still a reliable question? Maybe good alternative question?

Back Translation of the RUMUX-LITE into English

1. The opportunities of this system comply with my requirements
2. This system is easy to use

Appendix F: Confirmatory factor analysis of the RUS-BUS-11 model



Appendix G: Code from R

```
install.packages("readxl")      ### reads excel
install.packages("lavaan")      ### does LAtent VAriable ANalysis see
install.packages("lavaanPlot")  ### make plots
install.packages("dplyr")
install.packages("haven")
install.packages("ggpubr")
install.packages("semPlot")
install.packages("MVN")
install.packages("tidyverse")
install.packages("WriteXLS")
install.packages("ltm")
```

```
install.packages("outliers")
install.packages("EnvStats")
###############pull packages out of the library
library(readxl)
library(foreign)
library(lavaan)
library(lavaanPlot)
library(dplyr)
library(haven)
library(ggpubr)
library(knitr)
library(semPlot)
library(MVN)
library(tidyr)
library(tidyverse)
library(WriteXLS)
library(ltm)
library(outliers)
library(EnvStats)
###############turn off scientific notation
options(scipen = 999)
###############read in all data
BUS_CFA3 <- read.csv ("1.csv")
View(BUS_CFA3)
summary(BUS_CFA3)
###############Normality check
shapiro.test(BUS_CFA3$i2)
shapiro.test(BUS_CFA3$i3)
shapiro.test(BUS_CFA3$i4)
shapiro.test(BUS_CFA3$i5)
```

```
shapiro.test(BUS_CFA3$i6)

shapiro.test(BUS_CFA3$i7)

shapiro.test(BUS_CFA3$i8)

shapiro.test(BUS_CFA3$i9)

shapiro.test(BUS_CFA3$i10)

shapiro.test(BUS_CFA3$i11)

##############Model that we want to confirm

model1 <- 'fac1 =~ i1+i2

        fac2 =~ i3 + i4 + i5

        fac3 =~ i6 + i7 + i8 + i9

        fac4 =~ i10

        fac5 =~ i11'

###################CFA Model

fit <- cfa(model1, data = BUS_CFA3, estimator="MLR", mimic="Mplus")

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



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

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


##############Reliability ALL

##############BUS 11 reliability

alphaBUS11E3 <-data.frame(BUS_CFA3$i1,BUS_CFA3$i3,
BUS_CFA3$i3,BUS_CFA3$i4,BUS_CFA3$i5,BUS_CFA3$i6,BUS_CFA3$i7,BUS_CFA3$i
8,BUS_CFA3$i9, BUS_CFA3$i10, BUS_CFA3$i11)

##############F1= reliability

cronbach.alpha(alphaBUS11E3)

alphaF1E3<-data.frame(BUS_CFA3$i1,BUS_CFA3$i2)

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

```
#############F2= reliability
alphaF1E3<-data.frame(BUS_CFA3$i3,BUS_CFA3$i4,BUS_CFA3$i5)
cronbach.alpha(alphaF1E3, standardized = TRUE, CI = TRUE)

############F3= reliability
alphaF1E3<-data.frame(BUS_CFA3$i6,BUS_CFA3$i7,BUS_CFA3$i8,BUS_CFA3$i9)
cronbach.alpha(alphaF1E3, standardized = TRUE, CI = TRUE)

############F4= reliability
alphaF1E3<-data.frame(BUS_CFA3$i10)
cronbach.alpha(alphaF1E3, standardized = TRUE, CI = TRUE)

############F5= reliability
alphaF1E3<-data.frame(BUS_CFA3$i11)
cronbach.alpha(alphaF1E3, standardized = TRUE, CI = TRUE)

##########Correlation With Other scales: BUS correlates with UMXLITE Distribution
shapiro.test(BUS_CFA3$Utotal)
shapiro.test(BUS_CFA3$BUStotal)

cor.test(BUS_CFA3$BUStotal, BUS_CFA3$Utotal,use="pairwise.complete.obs", method =
"kendall")
##########correlation test for each factor
cor.test(BUS_CFA3$f1, BUS_CFA3$Utotal,use="pairwise.complete.obs", method =
"kendall")
cor.test(BUS_CFA3$f2, BUS_CFA3$Utotal,use="pairwise.complete.obs", method =
"kendall")

cor.test(BUS_CFA3$f3, BUS_CFA3$Utotal,use="pairwise.complete.obs", method =
"kendall")
```

```
cor.test(BUS_CFA3$f4, BUS_CFA3$Utotal,use="pairwise.complete.obs", method =
"kendall")


cor.test(BUS_CFA3$f5, BUS_CFA3$Utotal,use="pairwise.complete.obs", method =
"kendall")

shapiro.test(BUS_CFA3$BUStotal)

shapiro.test(BUS_CFA3$Utotal)

cor.test(BUS_CFA3$Utotal, BUS_CFA3$BUStotal)

##########Cronbach´s alpha of the RUMUX-LITE

alphaUMUXE3 <-data.frame(BUS_CFA3$u1,BUS_CFA3$u2)

cronbach.alpha(alphaUMUXE3)

##########simple linear regression between age and BUStotal

results <-  lm(BUStotal ~age, data=BUS_CFA3)

Summary(results)
```