Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?

Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?

Samuel Holmes
Ulster University
School of Communication & Media
Newtownabbey, UK
holmes-w@ulster.ac.uk

Anne Moorhead
Ulster University
School of Communication & Media/Institute of Nursing & Health Research
Newtownabbey, UK
a.moorhead@ulster.ac.uk

Raymond Bond
Ulster University
School of Computing
Newtownabbey, UK
rb.bond@ulster.ac.uk

Huiru Zheng
Ulster University
School of Computing
Newtownabbey, UK
h.zheng@ulster.ac.uk

Vivien Coates
Ulster University
Institute of Nursing & Health Research
Coleraine, UK
ve.coates@ulster.ac.uk

Michael McTear
Ulster University
School of Computing
Newtownabbey, UK
mf.mctear@ulster.ac.uk

ABSTRACT

Chatbots are becoming increasingly popular as a human-computer interface. The traditional best practices normally applied to User Experience (UX) design cannot easily be applied to chatbots, nor can conventional usability testing techniques guarantee accuracy. *WeightMentor* is a bespoke self-help motivational tool for weight loss maintenance. This study addresses the following four research questions: How usable is the *WeightMentor* chatbot, according to conventional usability methods?; To what extend will different conventional usability questionnaires correlate when evaluating chatbot usability?; And how do they correlate to a tailored chatbot usability survey score?; What is the optimum number of users required to identify chatbot usability issues?; How many task repetitions are required for a first-time chatbot users to reach optimum task performance (i.e. efficiency based on task completion times)? This paper describes the procedure for testing the *WeightMentor* chatbot, assesses correlation between typical usability testing metrics, and suggests that conventional wisdom on participant numbers for identifying usability issues may not apply to chatbots. The study design was a usability study. *WeightMentor* was tested using a pre-determined usability testing protocol, evaluating ease of task completion, unique usability errors and participant opinions on the chatbot (collected using usability questionnaires). *WeightMentor* usability scores were generally high, and correlation between questionnaires was strong. The optimum number of users for identifying chatbot usability errors was 26, which challenges previous research. Chatbot users reached optimum proficiency in tasks after just one repetition. Usability test outcomes confirm what is already known about chatbots - that they are highly usable (due to their simple interface and conversation-driven functionality) but conventional methods for assessing usability and user experience may not be as accurate when applied to chatbots.

ECCE '19, September 10–13, 2019, Belfast, UK

ACM ISBN 978-1-4503-7166-7-9/10/19.

CCS CONCEPTS

• **Human-centered computing** → **Usability testing**; • **Software and its engineering** → **Software testing and debugging**.

KEYWORDS

Usability Testing, Chatbots, Conversational UI, UX Testing

ACM Reference Format:
Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *ECCE '19: European Conference on Cognitive Ergonomics, September 10–13, 2019, Belfast, UK*. ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Chatbots are becoming increasingly popular as a human-computer interface. The year 2016 was described as "The rise of the chatbot" [20], and major companies including Microsoft, Google, Amazon and Apple have all developed and deployed their own "personal digital assistants" or "smart speakers" which are platforms for chatbots (also known as voicebots). Interacting with a chatbot is arguably more natural and intuitive given that it is like human-human interaction when compared to conventional methods for human-computer interaction. Moreover, given that chatbots integrate with popular social media platforms such as Facebook Messenger or Skype, users are not required to learn new unfamiliar interfaces or even download an app.

1.1 Chatbot UX Design

Cameron et al. (2018) suggested that the chatbot development life-cycle is different from traditional development life-cycles [7]. For example, where conventional user interface design may focus on user

interface prototyping, chatbot design instead focuses on conversation modelling, and the "interface" may be improved by analysing interaction logs to determine how best to improve conversation flow [7]. Best practices for User Experience (UX) design, such as Schneiderman's Eight Golden Rules [27] and Nielsen's Ten Usability Heuristics [18] cannot be easily applied to chatbots, which must instead replicate human conversation. Moore et al. (2017) identify three basic principles of conversation design, and four main conversation types, summarized in Tables 1 and 2 [17].

Table 1: Basic conversation design principles (Moore et al. 2017)

Principle	Description
Recipient Design	Tailoring conversation to match user's level of understanding
Minimization	Keeping interactions as short and simple as possible
Repair	Recovering from failures and helping understanding (e.g. repeating/rephrasing)

Table 2: Main conversation types (Moore et al. 2017)

Conversation Type	Description
Ordinary Conversation	Casual and unrestrained (e.g. small talk)
Service Conversation	Constrained by rules and roles (e.g. customer service agent/customer)
Teaching Conversation	Happens between teacher and student. Teacher probes to test knowledge
Counselling Conversation	Counselee leads the conversation, seeking advice from the counsellor

1.2 Testing Usability of Chatbots

Usability tests are generally conducted according to standard practices using standardized tools. However, in some cases it is necessary to modify test protocols to account for characteristics of participants or technologies. Gibson et al. (2016) described a usability test of an app designed for aiding reminiscence in people living with dementia. In this case it was necessary to favour usability testing metrics such as task completion over others such as concurrent think-aloud and usability questionnaires [12]. It may be necessary to similarly modify traditional usability testing methods when testing chatbots. UX24/7 describe four main issues that may be encountered by users of a traditional system such as a website. These four issues are Language, Branding, Functions and Information Retrieval [29]. If language used on a website is too complex, a user may struggle to understand. Chatbot users may face the same issue, as chatbots are conversation-based. Branding in websites and software is generally always visual, using recognizable graphics and colours. Chatbots, on the other hand, are conversation driven, thus the conversation and tone of voice need to reflect the brand. If functions of a website or system are poorly designed, this will reduce the usability of the site. In conversation-based systems, functions may be considered equal to conversations, and conversation flow, which, if poorly designed will also affect usability. Finally, information retrieval must be accurate. In a web-based system for example, a poorly designed search function may result in incorrect information being returned to the user, but in chatbot terms, this may happen if the chatbot incorrectly interprets what the user says or misunderstands their question [29].

A 2018 study by Nielsen-Norman group suggested that several aspects of chatbots should be tested in order to validate the UX [5]. These include interaction style (e.g. buttons and links vs. text entry), conversation flow, language and privacy. Nevertheless, it is evident that testing the usability of chatbots might require new methods beyond the conventional usability engineering instruments since chatbots offer a very different kind of human-computer interaction. Furthermore, given that usability validation of medical devices and healthcare software is often a requirement for FDA approval, measuring the usability of a healthcare focused chatbot is an important research topic.

1.3 Chatbots in Healthcare

Whilst the research is primitive, chatbots have been shown to be of use as "therapeutic" healthcare interventions or for at least augmenting traditional healthcare interventions. Barak et al. (2009) reported that the ability of chatbots to create empathy and react to emotions resulted in higher compliance with therapeutic treatments [3]. Healthcare focused chatbots facilitate increased user engagement and increased usability [10] and may also solve issues with text message-based systems, such as 24-hour availability and automated messages sounding less human [9]. In response to growing waiting lists and difficulties in accessing mental health services, Cameron et al. (2018) developed iHelpr, a self-help mental health chatbot. It is suggested that conversational interfaces may soon replace web pages and smartphone apps as the preferred means of conducting online tasks [8]. Chatbots are suggested to be of use in the area of mental health because they provide instant access to help and support and increased efficiency [6].

1.4 The *WeightMentor* Chatbot

In this paper, we describe a study that assessed the usability of a healthcare focused chatbot called *WeightMentor*, which we have developed at Ulster University. This chatbot is a bespoke self-help motivational tool for weight loss maintenance, with the purpose of supports weight loss maintenance by encouraging self-reporting, personalized feedback, and motivational dialogues [13]. Self-reporting, personalized feedback and motivation have been shown to be beneficial for weight loss maintenance in the short term [9] [11]. This paper involves the usability testing of the *WeightMentor* chatbot and uses this experiment as a case study to help answer several key research questions in this field.

2 RESEARCH QUESTIONS

(1) How usable is the *WeightMentor* Chatbot, according to conventional usability methods?
(2) To what extend will different conventional usability questionnaires correlate when evaluating chatbot usability? And how do they correlate to a tailored chatbot usability survey score?
(3) What is the optimum number of users required to identify chatbot usability issues?
(4) How many task repetitions are required for a first-time chatbot users to reach optimum task performance (i.e. efficiency based on task completion times)?

Question 3 is of interest as knowing how many subjects to recruit to identify most of the usability issues in a chatbot is important for planning usability studies in the UX industry. Moreover, it is also important given that studies have reported that most usability issues of traditional human-computer systems can be identified when recruiting 5-8 subjects [19]. However, this may not hold true for chatbot interfaces.

3 METHODS

3.1 Research Design

This usability test was an observational research design, testing the usability of the chatbot, and comparing a novel usability questionnaire specifically designed for chatbot testing with existing questionnaires. Ethical approval for this study was obtained from the School of Communication & Media Filter (Research Ethics) Committee, Ulster University in October 2019.

3.2 Chatbot Development

The *WeightMentor* chatbot was developed for Facebook Messenger, as Facebook is currently the most popular social media tool on smartphones and mobile devices [21] [22]. The chatbot conversation flow was designed using DialogFlow, which is a powerful framework integrating Google's machine learning and natural language understanding capabilities. The basic chatbot functionality was supplemented using a custom built NodeJS app hosted on Heroku.

3.3 Test Protocol

Figure 1 shows a photograph of a typical usability testing setup. The MOD1000 camera records the user's hand movements as they interact with the chatbot. The laptop runs software to capture video from the MOD1000 and audio as part of the concurrent think-aloud aspect of the usability test.

Usability tests were conducted as follows:

(1) Participant signed the consent form
(2) Participant was given access to *WeightMentor*
(3) Test coordinator read a briefing to the participant
(4) Participant completed pre-test (demographic) questionnaire
(5) Test coordinator read each task to the participant
(6) Participant confirmed their understanding of the task

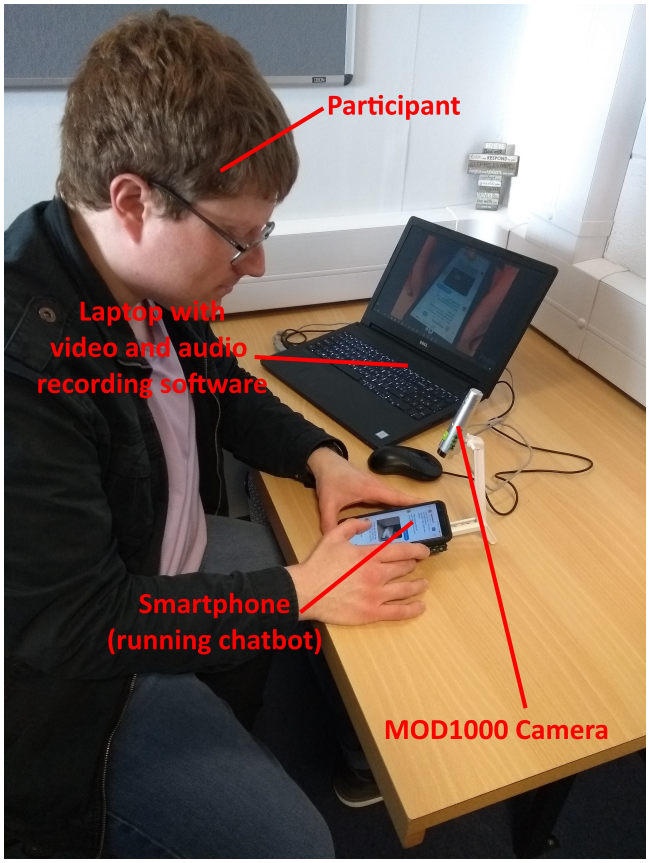

Figure 1: Typical usability testing setup.

(7) Participant answered pre-task Single Ease Question ("On a scale of 1 to 7, how easy do you think it will be to complete this task?")
(8) Participant was video and audio recorded completing task. Participant talked through what they could see on screen, and what they were trying to do during task (known as the concurrent think aloud protocol)
(9) Participant answered post-task Single Ease Question ("On a scale of 1 to 7, how easy was it to complete this task?")
(10) After all tasks, each participant completed Post-Test usability surveys including System Usability Scale (SUS) survey [4], User Experience Questionnaire (UEQ) [15] and a new Chatbot Usability Questionnaire (CUQ) that we developed for this study (refer to figure 2).

3.4 Usability Metrics

Baki Kocaballi et al. (2018) suggested that multiple metrics may be more appropriate for measuring chatbot usability [14]. Thus, three metrics were selected to evaluate the usability of *WeightMentor*: the SUS scores, UEQ metrics and our own CUQ score (although CUQ is yet to be validated but is has been designed to measure the usability of chatbots as opposed to measuring the general usability of human-computer systems). In addition, we measured task completion times,

usability issues and the differential between pre-task SEQ answers and post-task SEQ answers.

3.4.1 System Usability Scale (SUS). SUS was designed as a quick and easy means of assessing usability [4] and today is one of the most commonly used usability testing tools. SUS is comprised of ten validated statements, covering five positive aspects and five negative aspects of the system. Participants score each question out of five. Final scores are out of 100 and may be compared with the SUS benchmark (currently 68.0 representing an average score) or interpreted using several possible scales.

SUS scores can be grouped into percentile ranges [25]. Alternatively, a grade system may be used. Bangor et al. (2009) propose an adjective based scale, ranging from "Worst Imaginable" (SUS grades 0 - 25) to "Best Imaginable" (SUS grades over 84.1) [2]. The Acceptability Scale [1] ranges from "Not Acceptable" (SUS < 62.6) to "Acceptable" (SUS > 71.1). Finally, SUS scores may be linked with the Net Promoter Score (NPS), designed for measuring customer loyalty [23]. SUS scores greater than 78.8 may be classed as "Promoters", below 62.7 will be "Detractors", and scores in between are "Passive" [24].

The benchmark SUS score falls within the 41st - 59th percentile. It is grade C, "Good" on the adjective scale, "Marginal" on the acceptability scale, and NPC classification "Passive".

3.4.2 User Experience Questionnaire (UEQ). The UEQ serves as a means of comprehensively assessing the UX [15]. It is based on six scales, summarised in Table 3. Scales are measured using pairs of opposite adjectives to describe the system, with participants selecting their level of agreement with each. UEQ scores assess the extent to which the system meets expectations, but more usefully may be compared with a benchmark to determine how the system under test compares to other systems.

Table 3: UEQ Scales

Scale	What does it measure?
Attractiveness	Extent to which users "like" the system
Perspicuity	Ease of learning and becoming proficient in the system
Efficiency	Effort required to complete tasks, System reaction times
Dependability	Extent of user control, System predictability/security
Stimulation	How fun/exciting is it to use the system?
Novelty	Creativeness/Interest for users

3.4.3 Chatbot Usability Questionnaire (CUQ). The CUQ is based on the chatbot UX principles provided by the ALMA Chatbot Test tool [16], which assess personality, onboarding, navigation, understanding, responses, error handling and intelligence of a chatbot. The CUQ is designed to be comparable to SUS except it is bespoke for chatbots and includes 16 items. Participants' levels of agreement with sixteen statements relating to positive and negative aspects of the chatbot are ranked out of five, from "Strongly Disagree", to "Strongly Agree". Statements used in the CUQ are listed in Table 4.

Table 4: Statements used in the novel and bespoke but 'unvalidated' Chatbot Usability Questionnaire (CUQ)

Question Number	Question
1	The chatbot's personality was realistic and engaging
2	The chatbot seemed too robotic
3	The chatbot was welcoming during initial setup
4	The chatbot seemed very unfriendly
5	The chatbot explained its scope and purpose well
6	The chatbot gave no indication as to its purpose
7	The chatbot was easy to navigate
8	It would be easy to get confused when using the chatbot
9	The chatbot understood me well
10	The chatbot failed to recognise a lot of my inputs
11	Chatbot responses were useful, appropriate and informative
12	Chatbot responses were not relevant
13	The chatbot coped well with any errors or mistakes
14	The chatbot seemed unable to handle any errors
15	The chatbot was very easy to use
16	The chatbot was very complex

3.5 Questionnaire Analysis

3.5.1 SUS Calculation. The SUS score calculation spreadsheet [28] was used to calculate SUS scores out of 100. The formula for this calculation is shown in equation 1.

$$\overline{SUS} = \frac{1}{n} \sum_{i=1}^{n} norm. \sum_{j=1}^{m} \begin{cases} q_{i,j}-1, q_{i,j} mod 2 > 0 \\ 5-q_{i,j}, otherwise. \end{cases} \quad (1)$$

where n=number of subjects (questionnaires), m=10 (number of questions), $q_{i,j}$=individual score per question per participant, norm=2.5.

3.5.2 UEQ Calculation. The UEQ Data Analysis Tool [26] analyses questionnaire data and presents results graphically. By default, the UEQ does not generate a single score for each participant but instead provides six scores, one for each attribute. To facilitate correlation analysis using the UEQ, a âĂIJmean UEQ scoreâĂİ from the six scores was calculated for each participant. The mean UEQ score is the mean of the scores for all six scales of the UEQ, per participant.

3.5.3 CUQ Calculation. CUQ scores were calculated out of 160 using the formula in equation 2, and then normalized to give a score out of 100, to permit comparison with SUS.

$$CUQ = \left(\left(\sum_{n=1}^{m} 2n-1\right)-5\right) + \left(25-\left(\sum_{n=1}^{m} 2n\right)\right) \times 1.6 \quad (2)$$

where m = 16 (number of questions) and n = individual question score per participant.

3.6 Data Analysis & Presentation Tools

R Studio and Microsoft Excel were used for data analysis. Box plots were produced using R Studio, and bar/line plots were produced using excel. Statistical tests, such as correlation tests and t-tests were conducted in R Studio. Correlation analysis was conducted using Pearson correlation coefficient. Mean task completion times were compared for significance using paired Wilcoxon analysis. The p-value threshold for statistical significance was 5% (or 0.05).

4 RESULTS

4.1 *WeightMentor* Chatbot Usability

4.1.1 System Usability Scale. A total of 30 participants (healthy adults) were recruited and who evaluated the usability of the *WeightMentor* chatbot. A boxplot of *WeightMentor* SUS scores is shown in Figure 2. The mean *WeightMentor* SUS score was 84.83 ± 12.03 and the median was 86.25. The highest score was 100.0 and the lowest was 57.50. The mean *WeightMentor* score places the chatbot in the 96th-100th percentile range, equivalent to Grade A+, "Best Imaginable", "Acceptable" and NPS Class "Promoter" on the various scales discussed in methods. Hence, the chatbot has a high degree of usability according to traditional usability SUS scores. However, SUS distributions for benchmarking have not included usability scores from testing chatbots.

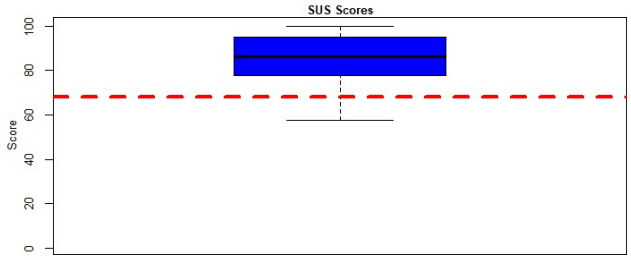

Figure 2: *WeightMentor* SUS scores (Benchmark of 68.0 is marked).

4.1.2 User Experience Questionnaire. The chatbot scored highly in all UEQ scales. Scores were well above benchmark and are presented graphically in Figure 3. Participant scores for each scale were all above +0.8, suggesting that in general participants were satisfied with the *WeightMentor* user experience.

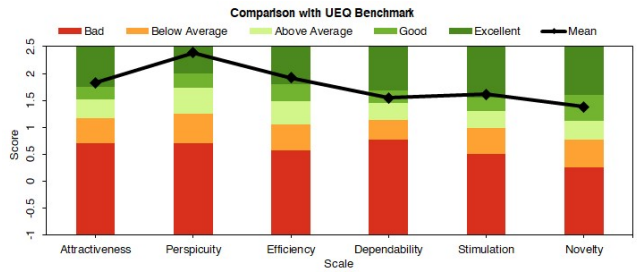

Figure 3: *WeightMentor* UEQ scores against benchmark.

4.1.3 Chatbot Usability Questionnaire. A box plot of *WeightMentor* CUQ scores is shown in Figure 4. The mean score was 76.20 ± 11.46 and the median was 76.5. The highest score was 100.0 and lowest was 48.4.

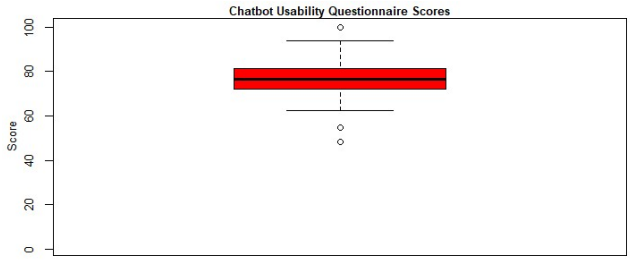

Figure 4: *WeightMentor* CUQ scores.

4.2 Usability Questionnaire Correlation

Correlations between each usability survey score and scatter plots are shown in Figure 5. All correlations are statistically significant since the p-values are all below 0.05 (5%). Multiple Regression was used to determine if CUQ score can be determined by SUS and Mean UEQ score. Results are shown in Table 5.

4.3 Chatbot Usability Issues

Thirty participants identified fifty-three usability issues. Usability issues were selected based on think-aloud data and feedback from participants in the Post-Test Survey. Usability issues identified per participant were listed in a spreadsheet, in the chronological order in which the tests were conducted. All usability issues identified by the first participant were treated as unique, and usability issues identified by subsequent participants were considered unique if they had not been identified by a previous participant. All unique usability issues identified at each test are plotted as a line graph in Figure 6.

To determine the best-case scenario, participants were re-sorted in a spreadsheet in descending order based on the number of usability issues identified per participant. Where more than one participant identified the same number of issues, these were sorted in chronological order of participants. Usability issues identified by the first participant were treated as unique and subsequent issues were counted only if they had not yet been identified by previous participants. Identified usability issues were plotted against number of participants as a line graph, shown in Figure 6. In this scenario, the first participant identifies the most usability issues, and the last participant identifies the least. This represents the best case (i.e. the least number of subjects required to identify almost all of the usability issues).

To determine the worst-case scenario, participants for best case scenario were sorted in reverse order, and usability issues were identified as for previous line graphs. The graph of this scenario is also shown in Figure 6, where the first participant identified the fewest usability issues, whilst the last participant identified the most usability issues.

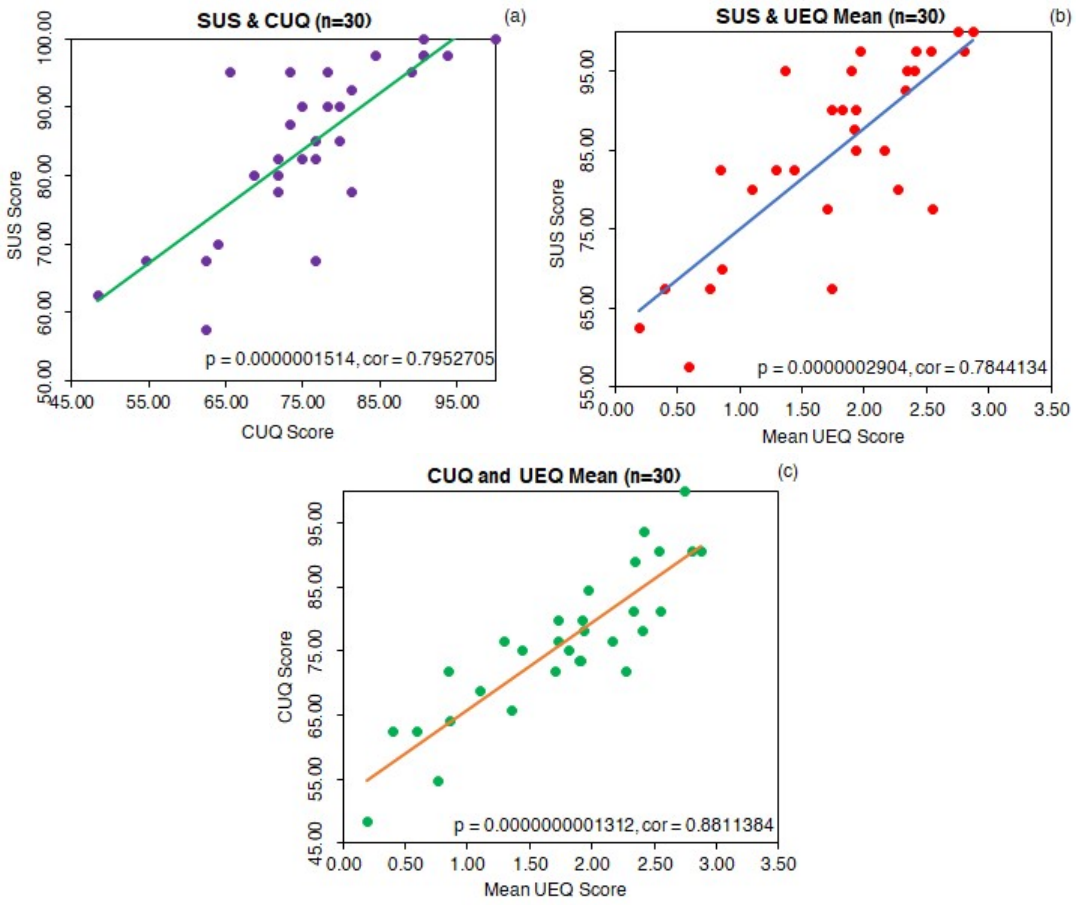

Figure 5: Scatter plots of (a) SUS and CUQ scores (b) SUS and UEQ mean scores (c) CUQ and UEQ mean scores

Table 5: Multiple regression results, where SUS and UEQ are independent variables and CUQ is the dependent variable

Coefficients	Estimate	Std. Error	t-value	p-value
Intercept	36.0128	8.5217	4.226	0.000243
SUS	0.2578	0.1307	1.973	0.058854
UEQ	10.3694	2.1264	4.876	0.0000425
Residual Standard Error	5.253 on 27 degrees of freedom			
Multiple R-Squared	0.8046	Adjusted R-Squared		0.7901
F-statistic	55.58 on 2 and 27 DF		p-value	0.0000000002681

4.4 Chatbot Task Completion

During the usability tests, participantsâĂŹ interactions with the chatbot were video recorded. Task completion times were calculated for each individual participant, along with the mean time overall. A benchmark task completion time was established by recording the performance of the developer (SH). Two tasks (task 2 and task 3) were repeated four times by participants to determine if task completion times improved with each repetition. Bar plots of task completion times against the benchmark are shown in Figure 7. The p-values for repeated tasks are shown in Table 6.

5 DISCUSSION

WeightMentor mean SUS score places the chatbot at the top of the various scales discussed in section 3.3.1 above. This suggests that in comparison to conventional systems, *WeightMentor* is highly usable. Similarly, *WeightMentor* UEQ scores were favourable when compared to UEQ benchmark. However, SUS and UEQ benchmark scores are derived from tests of conventional systems such as websites, therefore do not include any chatbot scores. Thus, while *WeightMentor* may score highly using these metrics, it is impossible to determine the accuracy of this score relative to other chatbots.

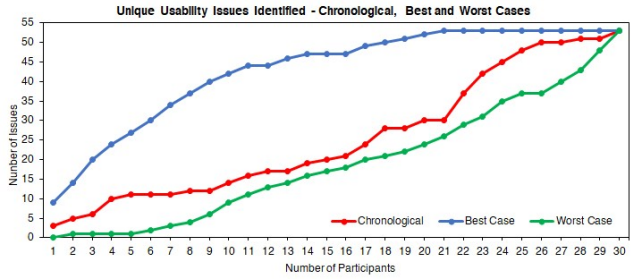

Figure 6: Usability issues identified against number of participants.

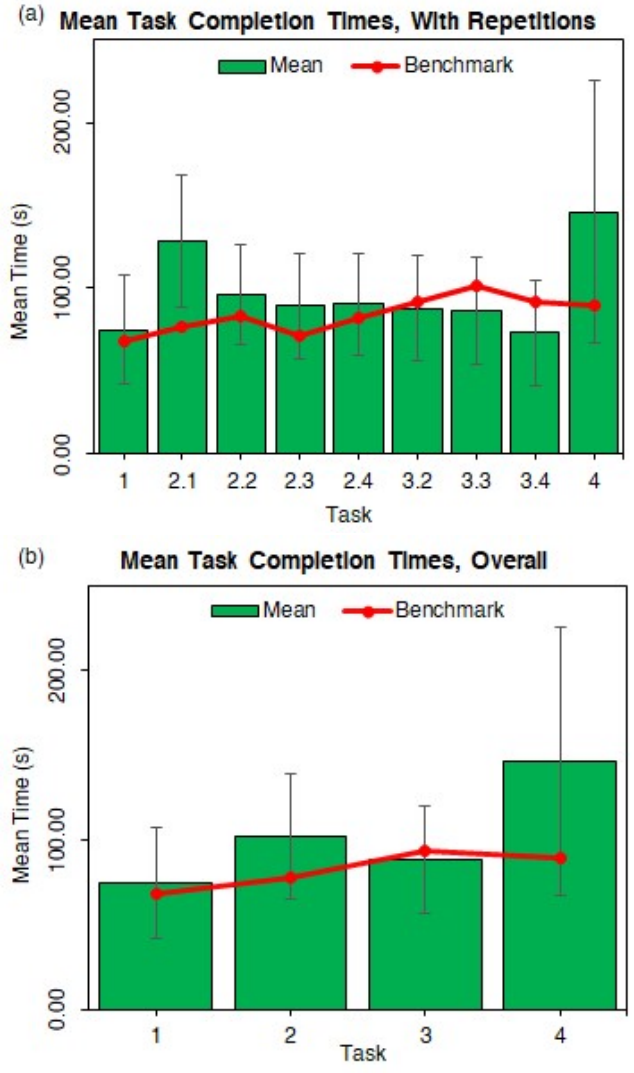

Figure 7: Task completion times against benchmark (a) per task with repetitions (b) per task overall

Correlation between the three main questionnaires was generally strong and was highest between the CUQ and UEQ. Multiple Regression results suggest that 80% of the variance in CUQ score can be explained by SUS and UEQ. Thus 20% of the variance in CUQ score is explained by other factors i.e. CUQ is perhaps measuring constructs more closely related to chatbots that is not being measured by SUS and UEQ. The p-value for SUS is greater than 0.05 (5%) thus it is not statistically significant in predicting CUQ scores, and the p-value for UEQ is less than 0.05 hence this is statistically significant in predicting CUQ score. This indicates that UEQ has a greater influence on CUQ score than SUS, which suggests that UEQ and CUQ are measuring similar aspects of chatbot usability and UX. This finding can be rationalized given that UEQ is more complex when compared to SUS. The implication of these findings is that SUS may be less effective for measuring chatbot usability on its own as it may not measure aspects of UX that make chatbots particularly usable.

Table 6: Task completion time p-values

Comparison		p-value
Task 2	Attempt 1 & Attempt 2	0.002621
	Attempt 2 & Attempt 3	0.3457
	Attempt 3 & Attempt 4	0.8284
Task 3	Attempt 1 & Attempt 2	0.06736
	Attempt 2 & Attempt 3	0.9249
	Attempt 3 & Attempt 4	0.1031

Analysis of identified usability issues suggested that in the best-case scenario, the optimum number of users is 21 (since the function in figure 6 plateaus thereafter). In the worst-case scenario, the optimum number of users is 30. In chronological order the optimum number was 26 - 29 users. The mean number of users required to identify most of the usability issues in the chatbot is 26. Nielsen & Landauer (1993) determined that 80% of unique usability issues may be captured by no more than 5 to 8 users, however this research concerned conventional systems, not chatbots, and it may be the case that the nature of chatbots makes it more difficult to identify usability issues with a smaller number of participants [19].

In general, task completion times did improve with each repetition of a task, however while the time difference was statistically significant between the first attempt and the second attempt at task 2, it was not statistically significant between the second and third and third and fourth attempts. Similarly, time differences were not statistically significant between any of the attempts at task 3, which may be because task 3 was very similar in procedure to task 2. This suggests that it may be possible for users to become proficient with a new chatbot very quickly, owing to their simplicity and ease of use.

6 CONCLUSION

Chatbot UX design and usability testing may require nontraditional methods, and multiple metrics are likely to provide a more comprehensive picture of chatbot usability. The *WeightMentor* chatbot scored highly on both SUS and UEQ scales, and correlation analyses suggest that correlation is stronger between CUQ and UEQ than CUQ and SUS, and validation of the CUQ will increase its

effectiveness as a usability analysis tool. Given that the variance of CUQ scores are not completely explained by SUS and UEQ, there is an argument that these traditional usability testing surveys do not evaluate all aspects of a chatbot interface. This study also suggests that approximately 26 subjects are required to identify almost all the usability issues in the chatbot which challenges previous research. Whilst primitive, this work also suggests that users become optimal after just one attempt of a task when using a chatbot. This could be explained by the fact that chatbots can be less complex in that they lack visual hierarchy and complexity as seen in normally graphical user interfaces. This study suggests that conventional usability metrics may not be best suited to assessing chatbot usability, and that metrics will be most effective if they measure aspects of usability that are more closely related to chatbots. Chatbot usability testing may also potentially require a greater number of users than suggested by previous research, in order to maximize capture of usability issues. Finally, this study suggests that users can potentially reach optimum proficiency with chatbots very quickly.

ACKNOWLEDGMENTS

This research study is part of a PhD research project at the Ulster University, funded by the Department for the Economy, Northern Ireland.

REFERENCES

[1] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human Computer Interaction* 24, 6 (07/29 2008), 574–594. https://doi.org/10.1080/10447310802205776 doi: 10.1080/10447310802205776.

[2] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4, 3 (2009), 114–123.

[3] Azy Barak, Britt Klein, and Judith G. Proudfoot. 2009. Defining Internet-Supported Therapeutic Interventions. *Annals of Behavioral Medicine* 38, 1 (08/01 2009), 4–17. https://doi.org/10.1007/s12160-009-9130-7 ID: Barak2009.

[4] John Brooke. 1996. *SUS: A 'quick and dirty' usability scale.* CRC Press, 189–194.

[5] Raluca Budiu. 2018. The User Experience of Chatbots.

[6] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a Chatbot for Digital Counselling. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference (HCI '17)*. BCS Learning & Development Ltd, Swindon, UK, 24:1–24:7. https://doi.org/10.14236/ewic/HCI2017.24

[7] Gillian Cameron, David Cameron, Gavin Megaw, RR Bond, Maurice Mulvenna, Siobhan O'Neill, C. Armour, and Michael McTear. 2018. *Back to the Future: Lessons from Knowledge Engineering Methodologies for Chatbot Design and Development.* https://doi.org/10.14236/ewic/HCI2018.153

[8] Gillian Cameron, David Cameron, Gavin Megaw, RR Bond, Maurice Mulvenna, Siobhan O'Neill, C. Armour, and Michael McTear. 2018. *Best practices for designing chatbots in mental healthcare âĂŞ A case study on iHelpr.* https://doi.org/10.14236/ewic/HCI2018.129"

[9] E. L. Donaldson, S. Fallows, and M. Morris. 2014. A text message based weight management intervention for overweight adults. *Journal of Human Nutrition & Dietetics* 27, Suppl 2 (Apr 2014), 90–97. http://ovidsp.ovid.com/athens/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=23738786

[10] AH Fadil and S. Gabrielli. 2017. Addressing Challenges in Promoting Healthy Lifestyles: The AI-Chatbot Approach.

[11] Brianna S. Fjeldsoe, Ana D. Goode, Philayrath Phongsavan, Adrian Bauman, Genevieve Maher, Elisabeth Winkler, and Elizabeth G. Eakin. 2016. Evaluating the Maintenance of Lifestyle Changes in a Randomized Controlled Trial of the 'Get Healthy, Stay Healthy' Program. *JMIR MHealth and UHealth* 4, 2 (2016), e42. http://ovidsp.ovid.com/athens/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=prem&AN=27166643

[12] Aideen Gibson, Claire McCauley, Maurice Mulvenna, Assumpta Ryan, Liz Laird, Kevin Curran, Brendan Bunting, Finola Ferry, and Raymond Bond. 2016. Assessing Usability Testing for People Living with Dementia. In *Proceedings of the 4th Workshop on ICTs for Improving Patients Rehabilitation Research Techniques (REHAB '16)*. ACM, New York, NY, USA, 25–31. https://doi.org/10.1145/3051488.3051492

[13] Samuel Holmes, Anne Moorhead, RR Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2018. A New Automated Chatbot for Weight Loss Maintenance. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI 2018)*. https://doi.org/10.14236/ewic/HCI2018.103"

[14] A. Baki Kocaballi, L. Laranjo, and E. Coiera. 2018. Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI 2018)*.

[15] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB*, Vol. 5298. 63–76. https://doi.org/10.1007/978-3-540-89350-9_6

[16] J. MartÃŋn, C. MuÃśoz-Romero, and N. ÃĄbalos. 2017. chatbottest - Improve your chatbot's design.

[17] Robert J. Moore, Raphael Arar, Guang-Jie Ren, and Margaret H. Szymanski. 2017. Conversational UX Design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 492–497. https://doi.org/10.1145/3027063.3027077

[18] J. Nielsen. 1994. 10 Heuristics for User Interface design.

[19] Jakob Nielsen and Thomas K. Landauer. 1993. A Mathematical Model of the Finding of Usability Problems. In *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems (INTERCHI '93)*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 206–213. http://dl.acm.org/citation.cfm?id=164632.164904

[20] Wharton University of Pennsylvania. 2016. The Rise of the Chatbots: Is It Time to Embrace Them?

[21] Ofcom. 2017. *The UK Communications Market: Telecoms & Networks.* Technical Report. Ofcom.

[22] Kelly Pedotto, Vivey Chen, and JP McElyea. 2017. The 2017 US Mobile App Report.

[23] Fred F. Reichheld. 2003. The One Number You Need to Grow.

[24] Jeff Sauro. 2012. Predicting Net Promoter scores from System Usability Scale Scores.

[25] Jeff Sauro. 2018. 5 ways to interpret a SUS score.

[26] Martin Schrepp. 2019. User Experience Questionnaire Data Analysis Tool.

[27] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, and N. Elmqvist. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (sixth edition ed.). Pearson.

[28] T. Tullis and B. Albert. 2018. Measuring the User Experience: SUS Calculation Tool.

[29] UX24/7. 2018. Chatbots: Usability Testing & Chatbots.