Usability testing and trust analysis of a mental health and wellbeing chatbot

Kyle Boyd
Ulster University
School of Art
Belfast, UK
ka.boyd@ulster.ac.uk

Courtney Potts
Ulster University
School of Computing
Belfast, UK
c.potts@ulster.ac.uk

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

Maurice D Mulvenna
Ulster University
School of Computing
Belfast, UK
md.mulvenna@ulster.ac.uk

Thomas Broderick
Munster Technological University
Department of Sport, Leisure and Childhood Studies
Cork, Ireland
thomas.broderick@mtu.ie

Con Burns
Munster Technological University
Department of Sport, Leisure and Childhood Studies
Cork, Ireland
con.burns@mtu.ie

Andrea Bickerdike
Munster Technological University
Department of Sport, Leisure and Childhood Studies
Cork, Ireland
andrea.bickerdike@mtu.ie

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

Catrine Kostenius
Luleå University of Technology
Department of Health Sciences
Luleå, Sweden
catrine.kostenius@ltu.se

Alex Vakaloudis
Munster Technological University
Nimbus Research Centre
Cork, Ireland
alex.vakaloudis@mtu.ie

Indika Dhanapala
Munster Technological University
Nimbus Research Centre
Cork, Ireland
indika.dhanapala@mtu.ie

Edel Ennis
Ulster University
School of Psychology
Coleraine, UK
e.ennis@ulster.ac.uk

Frederick Booth
Ulster University
School of Computing
Belfast, UK
F.Booth@ulster.ac.uk

ABSTRACT

Mental health chatbots are particularly useful for those who are isolated and may have difficulty attending services or for those who are reluctant to speak to a professional. In this study, the usability and trust of a chatbot known as 'ChatPal' has been assessed. ChatPal has been developed by an interdisciplinary team encompassing health service providers, local authorities, charities and universities to promote positive mental wellbeing among individuals in rural areas across Europe. This study employed a usability test protocol to recruit representative users to complete a set of tasks using the ChatPal chatbot. Usability issues were assessed along with trust and users' satisfaction on the System Usability Scale and the Chatbot Usability Questionnaire. The study shows the usability issues and trust with a mental health chatbot and highlights recommendations for improvement.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Health care information systems.*

KEYWORDS

Chatbots, Conversational user interfaces, User experience, Digital interventions, Apps, Mental health, Wellbeing

ACM Reference Format:
Kyle Boyd, Courtney Potts, Raymond Bond, Maurice D Mulvenna, Thomas Broderick, Con Burns, Andrea Bickerdike, Michael F McTear, Catrine Kostenius, Alex Vakaloudis, Indika Dhanapala, Edel Ennis, and Frederick Booth. 2022. Usability testing and trust analysis of a mental health and wellbeing chatbot. In *33rd European Conference on Cognitive Ergonomics (ECCE*

ECCE 2022, October 4–7, 2022, Kaiserslautern, Germany

ACM ISBN 978-1-4503-9808-4/22/10.
https://doi.org/10.1145/3552327.3552348

2022), October 4–7, 2022, Kaiserslautern, Germany. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3552327.3552348

1 INTRODUCTION

Chatbots, also known as conversational user interfaces, are tools that use machine learning and artificial intelligence to simulate human communication, either through voice or text communication [7] [17]. Voice-based chatbots are ubiquitous in contemporary society, embedded within mobile devices, computers and smart speakers such as 'Amazon Alexa' or 'Siri'. Text-based chatbots are widely available across platforms such as Messenger or Slack, the web or mobile devices. Chatbots typically use natural language processing which has been made possible through advancements in computing technology to enhance user experience and optimise personalised mental health care [2].

The promotion of mental health is one area within which chatbots can have a positive impact [28]. Given the increase in psychological distress resulting from the COVID-19 pandemic [22] [32] and the already overstretched resources of many healthcare providers with limited provisions available, digital interventions such as chatbots can be seen as a way to provide alternative support [27]. Mental health chatbots can be used in a blended approach to augment face-to-face services or potentially allow people to manage their own mental wellbeing without requiring external services.

1.1 Mental Health

Good mental health is a critical part of an individual's wellbeing and sets the foundation for a happy, fulfilled and productive life [9]. Mental ill health can impact people from all ages, backgrounds, and genders and affects about 84 million people across the EU countries [21]. WHO defines mental health as "a state of wellbeing in which the individual realises his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community" [30]. The chatbot assessed in this study focuses on positive psychology, one of the key attributes of mental health promotion [31]. Mental health is an important aspect in rural life but due to geographical isolation from traditional mental health services in many remote areas [29], there exists a particular rationale for digital services to promote self-care and wellbeing at the point of need as well as 24/7 [10]. Traditional one-to-one services support people with chronic mental illness as well as mild-to-moderate mental illness but this is expensive, and resource limited. One-to-one intervention support requires significant travel for clients living in rural areas; hence accessibility to traditional treatments is a particular concern. Additionally, citizens may feel embarrassed when setting up appointments with a support person due to the lack of anonymity.

1.2 Previous work

One of the most popular mental health chatbots is Woebot [6] which acts as a chatbot therapist, utilising cognitive behavioural therapy. The results of a randomised controlled trial found that Woebot significantly reduced symptoms of depression in two weeks [6]. Wysa is another popular mental health chatbot [13]. A study with Wysa looked at the effectiveness and engagement levels of users with self-reported symptoms of depression. The results showed promise given that average mood of those that used the chatbot extensively was significantly better than those that were less engaged [13]. CARO is a chatbot capable of having empathetic conversations and providing medical advice [11]. CARO can sense the conversational context, intent, and associated emotions. iHelpr is a mental health chatbot developed to provide self-assessment and guidelines for stress, anxiety, depression, sleep, and self-esteem [5]. Tess is another chatbot designed to provide support for treating anxiety and depression, shown to be a feasible option for offering support for those living with depression [8].

The current literature ([6], [13], [11], [5], [8]) demonstrates that mental health chatbots can be useful and helpful to support users' wellbeing. They are non-intrusive and can alleviate many of the barriers faced by users compared to a web page or mobile app. Participants like the anonymity of using chatbots rather than discussing feelings to a health professional [14]. That said, chatbots do have their limitations as they are not human and do not always understand the intricacies of human speech. This means that chatbots can provide nonsensical responses which could potentially be harmful when a person is in mental distress. That is why usability testing is an important aspect during a design sprint. The aim of this research is to provide some general recommendations for chatbot designers, usability, and trust researchers.

1.3 ChatPal

ChatPal is an NPA (Northern Periphery and Arctic Programme) project that involves developing a mental health chatbot to support individuals wellbeing. The chatbot allows people to look after their own mental wellbeing and aims to prevent the development of serious mental health problems. At the beginning of the project, mental healthcare professionals were surveyed to explore attitudes toward healthcare chatbots [27]. Needs analysis workshops took place across all partner regions with the general public, mental healthcare professionals and mental health service users to gather user needs which were turned into requirements for the chatbot [24]. Content for the ChatPal chatbot was then developed based on use cases that mental health professionals would endorse, and the user requirements identified in stakeholder workshops. The chatbot is centred around positive psychology, and users can converse with the app in multiple languages. However, this study focuses on usability testing of the English language version of the app.

The ChatPal chatbot architecture utilises the PhoneGap framework [23] for the front end and the Rasa framework [25] as the back end, with communication via HTTP requests/ responses. Upon receiving user inputs from the ChatPal app, the Rasa Natural Language Understanding (NLU) unit extracts user intentions and relevant metadata out of the input, which are known as intents and entities within the Rasa framework. Once the intents and entities are identified, corresponding responses are determined by the Rasa core. Rasa supports various responses including text messages, buttons, images and reminders. Rasa also allows custom actions which are executed separately in an action server which is run with the help of the Rasa SDK. The SQLAlchemy python library, which uses Postgres SQL queries, was used as the custom actions require database access.

The prototype version of ChatPal in English was released on android (Google Play store) and it can also be accessed on a web browser (Figure 1).

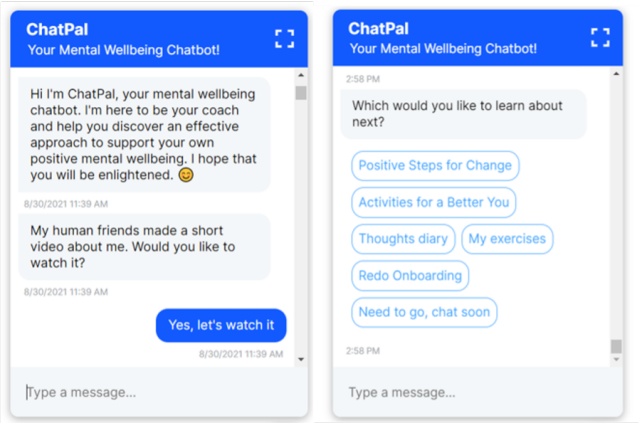

Figure 1: Screenshots of the ChatPal chatbot app interface. Left - initial greeting from the chatbot (grey boxes) with user selected response (blue box). Right - chatbot main menu with six options for user to select from.

ChatPal is available 24/7 and allows users to receive wellbeing support using text-based conversation. The chatbot is intuitive and uses language similar to everyday conversations allowing it to be adopted by those with poor computer literacy. The majority of chatbot conversations allow the user to select pre-defined text responses (Figure 1), with limited free text input. Within the chatbot, users can complete various activities, for example mindfulness techniques, breathing exercises and goal setting. They can also track their moods over time, store diary entries and receive information on various aspects of mental health. The present study seeks to assess the usability and trust of the ChatPal chatbot.

1.4 Research questions

This study examined the following research questions: What features of a mental health chatbot do users find acceptable and usable? How do people rate a mental health chatbot in terms of usability and trust? What recommendations can be made to optimise the usability and trust in a mental health chatbot?

2 METHODOLOGY

2.1 Research design

Usability testing was carried out with the ChatPal chatbot. Research design refers to the process of evaluating a product or service by testing it with representative users [20]. The goal of the test is to identify any usability problems, collect qualitative and quantitative data and determine the participant's satisfaction with the product [19]. One popular method is the concurrent think aloud protocol (TAP) [19]. In this method, the participant undertakes a series of tasks while thinking aloud in real-time during the task attempt. Some researchers prefer to use the retrospective think-aloud approach since people do not normally think-aloud during real world interactions with technologies and doing so in a lab-environment attenuates the fidelity of the test and makes it less natural. Some work on the usability of health chatbots has taken place ([5],[12], [3]). The present study utilised the TAP and retrospective think-aloud approaches.

2.2 Protocol

Ten participants were recruited to the usability study as outlined in Section 2.3. Due to the pandemic, usability testing took place online via conference calls using Zoom. The participants screen and audio was recorded during the session, to allow the recordings to be transcribed afterwards. Two observers fascilitated the meetings, explaining the study protocol and noting observations as participants were using the chatbot. Participants were asked to open the chatbot on their device, share their screen and complete a series of five tasks using the ChatPal chatbot. The tasks included:

(1) Go through the onboarding and consent sections. During onboarding, the chatbot provides a brief introduction explaining its purpose and asks for the user's first name/ nickname, age range, gender, and country.
(2) Go through an educational module. Participants were asked to go through a dialogue on sleep, in which they answered questions about their sleep routine and the chatbot shared advice.
(3) Log a mood and visualise graphs of mood. Participants select their mood by clicking on pictures (emojis). In the chatbot these are tracked over time, so people can view their overall moods or mood across days of the week.
(4) Log a gratitude diary entry. The chatbot explains the concept of gratitude and asks the user to input things they are grateful for. These are stored in the 'gratitude diary'.
(5) Take time to do some autonomous browsing. Participants were asked to browse the chatbot for 5-10 minutes and go through content of their choosing.

The intention was to have a set of common tasks that would be carried out while using the chatbot with the aim of highlighting design inconsistencies and usability problems within the user interface and content areas. After the five tasks, participants completed a post-test survey on Qualtrics which contained the System Usability Scale (SUS) [4] and the Chatbot Usability Questionnaire (CUQ) [12] and five questions on participant views towards trusting the chatbot to give advice and store information (mood logs, gratitude statements, and personal information). The SUS, which is used to measure usability, consists of a 10-item questionnaire with five response options from strongly agree to strongly disagree. While the SUS has been successful on its own, it may not be the best option for conversation-driven systems that do not necessarily conform to conventional design and testing principles. Thus, chatbot usability testing may require a different approach and so the CUQ was used alongside the SUS. The CUQ has been designed specifically to test the usability of chatbots at the post-test evaluation phase [12]. Combining these metrics will give us a more comprehensive assessment of the ChatPal chatbot.

2.3 Participants

Ethical approval was obtained from Ulster University Research Ethics Committee. Participants were recruited from Action Mental Health (a charity offering a myriad of mental health services in Northern Ireland) and Ulster University to take part in the study. Both groups were selected as ChatPal is targeted at promoting well-being in the general population and complementing existing mental health services, rather than treating mental ill health. Inclusion criteria for mental health clients included those who self-declared that they consider themselves to live in a rural area; had a history of mild-moderate anxiety and/or depression; were availing only of Action Mental Health services at the time and for the previous 6 weeks; were 18 years old or older; and had access to the internet and a device to use the app (computer or tablet or phone). Inclusion criteria for university staff and students included those who self-declared that they consider themselves to live in a rural area; never had a mental health diagnosis or had suicidal thoughts and behaviours in the previous year; were 18 years old or older and had access to the internet and a device to use the app (computer or tablet or phone).

The participants were given an information sheet to provide an opportunity to review the study and ask any questions before agreeing to take part. Written informed consent was obtained before commencing the study, and a distress protocol was put in place in the event that participants felt distress during the interview. Power analysis was not used to model the number of participants, as sample sizes of between 5 and 15 are deemed appropriate for usability testing, with the 5 yielding 80% of usability issues [20]. The tests took place in the participants' own homes in Northern Ireland and were recorded using Zoom (screen recording and audio only) and the recordings transcribed. A total of 10 participants completed the usability testing, 5 male and 5 female. Age range was between 21-48 (average 39). Of the 10 participants, 3 were university students and 7 were mental health service users. Computer literacy scores ranged from 2-5 (1-lowest, 5-highest), with an average of 3.6. Only 2 participants used the chatbot on their iPhone browser, while the rest opened ChatPal on their computer web browser.

2.4 Data analysis

Once all the recordings had been completed the data was collated and analysed using the programming language R and R Studio. Task completion times and the proportion of users that successfully completed each task were computed. Usability issues were noted from the written transcripts and rated for severity according to the Nielson Norman severity rating flow chart [18]. Total scores and summary statistics for SUS and CUQ were calculated for each participant. Pearson correlation coefficient was used to measure the relationship between numeric variables (age, computer literacy, SUS, CUQ and trust). Trust scores were calculated for each of the five questions (where 1=very low level of trust - 5 = very high level of trust) and for each participant by summing the responses to all questions. A Kruskal-Wallis test was used to assess if the responses were significantly different across the questions on trust.

3 RESULTS

3.1 Tasks and problems identified

Overall, at least 70% of users completed all tasks successfully (Figure 2). Generally, participants spent the longest on Task 3 (logging and visualising moods) and Task 5 (freely browsing the chatbot) (Figure 2). All participants easily completed Tasks 1 and 2 but had some difficulties or problems when trying to complete the other tasks (Table 1).

Most of the errors identified occurred just after Task 2 was completed (Table 1). While the chatbot was in use, the facilitators recorded some usability issues that the participants did not pick up on or mention. Some minor grammar errors were noted in the chatbot. Inconsistent text sizes were identified throughout all dialogues when viewing the chatbot on an iPhone browser.

Task	Proportion of users that successfully completed the task (%)	No. of usability issues
T1	100	2
T2	100	6
T3	80	2
T4	70	1
T5	70	2

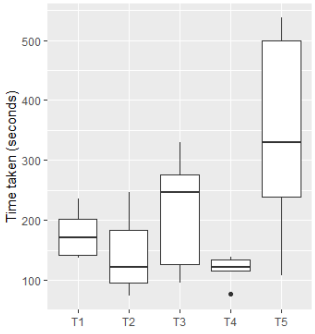

Figure 2: Completion and issues noted for tasks. T1 - onboarding. T2 - go through dialogue on sleep. T3 - log mood and visualise graphs of mood. T4 - log a gratitude diary entry. T5 - autonomous browsing

At the end of a dialogue, the chatbot asks the user if they would like to continue or go back to the main menu. After selecting an option, the user is asked to rate the conversation they just had with the chatbot before continuing or going to the menu (thumbs up - thumbs down). The facilitators felt it would be better for the conversational rating to appear at the end of the conversation, before asking the user what they would like to do next. This would improve the user experience as participants asked if they had to tap the menu button again, even though they had already selected it.

3.2 Usability issues

Participants and facilitators provided feedback which identified 14 usability issues with ChatPal, mostly specific to the ChatPal chatbot including two software bugs. The usability issues were rated by severity. Software bugs were imperative to fix, while most of the issues identified were high-medium priority to fix (n=8), and the rest were cosmetic problems. Participants suggested adding a video or text at the very beginning explaining what content is available in the chatbot and where to find it; and varying chatbot replies, as getting the same answers repeatedly would deter users from engaging with the chatbot.

Table 1: Issues identified by participants (P) during tasks.

Task and description	Success	Time required to complete task. Mean (SD)	Notes/ observations
1: Onboarding and consent	All participants easily completed	176.4 (37.7)	P1, P6 and P10 thought that the onboarding text appeared too fast. For P2 and P5, an error message appeared instead of the menu.
2: Go through educational module (sleep)	All participants easily completed	141.7 (60.5)	P1 felt large blocks of text were difficult to read and comprehend. For P2, the chatbot launched into a conversation before they could select any options from the menu. P2 felt the time stamps under replies were unnecessary and took up too much space, meaning they had to scroll back to read the full conversation. P3 felt the text appeared too fast, and that the grey text bubble might be hard for people with visual impairments to read. P3 and P7 thought the conversational rating (pictures showing thumbs up - thumbs down) was somewhat confusing. After completing this task, five of the participants experienced a technical error, where the chatbot got stuck in a loop repeating the previous conversation on sleep.
3: Log mood and visualise mood graphs	3 unable to complete, 5 completed with help, 2 easily completed	210.9 (93.4)	While trying to complete this task, P2, P5 and P10 got stuck in a loop of the sleep dialogue. P3 found the font on both mood graphs was too small to read. P7 completed mood logging but chose not to view graphs showing mood over time.
4: Log a gratitude diary entry	2 unable to complete, 2 completed with help, 6 easily completed	119.4 (20.9)	P9 wanted to write multiple gratitude entries, however the chatbot allowed only one entry at a time. P10 was unable complete this task due to a technical error
5: Freely browse chatbot	3 unable to complete, 1 completed with help, 7 easily completed	345.2 (174.7)	P1 got an unexpected error message while browsing the chatbot. P3 was confused by a menu titled 'my exercises', expecting it to have psychical exercises (for example, stretches) rather than previous activities completed in the chatbot. P10 found that most of chatbot dialogue appeared too quickly. During one of the conversations, a cloud emoji was presented as a button but the participant found this confusing and unsure if they should click this or not. P10 desired an option to go back to the main menu instead of going through the rest of the conversation in the chatbot.

3.3 System Usability Scale (SUS), Chatbot Usability Questionnaire (CUQ) and trust

The benchmark SUS score is 68 [15], where a score above a 68 would be considered above average and anything below 68 is below average. Four participants gave ChatPal a score below 68, while the rest scored above average (Figure 3). SUS scores ranged from 37.5-85 (Figure 3), with a mean of 64.3 and standard deviation (SD) = 16.8. There was a strong positive correlation between SUS and computer literacy (r=0.65, p=0.043), suggesting people with high computer literacy are likely to find the chatbot easier to use than those with low literacy. The CUQ is scored out of 100 so it can be used alongside and compared to SUS, which is also scored out of 100. There was a strong positive correlation between SUS and CUQ scores (r=0.69, p=0.028). The CUQ scores ranged from 54.7 - 76.6 (Figure 3), with a mean of 65 and SD=6.7. The CUQ and SUS scores indicate that the ChatPal chatbot as it currently stands is just below average in terms of usability. However, this is likely due to the obvious software bugs that were present which could be easily addressed by a technical team. There was no association between SUS or CUQ scores and age.

Overall trust in ChatPal was average to high (Figure 4). A larger percentage of participants were likely to trust ChatPal for storing gratitude statements, mood logs and giving advice than for storing personal information and stories (Figure 4). There were no significant differences in responses to the trust questions (Kruskal-Wallis chi-squared= 5.62, p=0.23, df=4). However, there was a strong positive correlation between participant's trust scores and CUQ (r=0.78, p=0.0076), indicating that those who rated the chatbot high in terms of usability were more likely to trust it. There was no association between trust and age, nor trust and computer literacy.

4 DISCUSSION

The aim of this study was to identify features that users found acceptable and usable, how they rate a mental health chatbot in terms of usability and trust and provide general recommendations for others designing mental health chatbots. This study benefited from

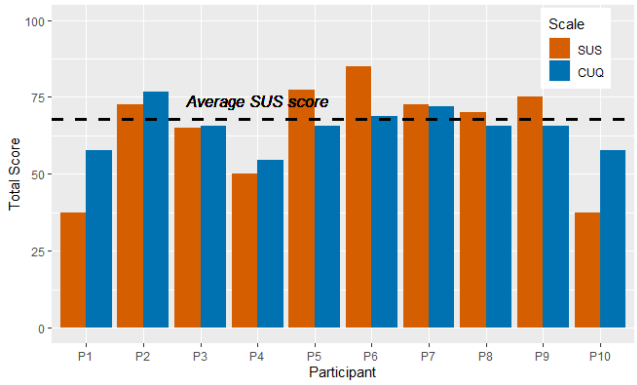

Figure 3: SUS and CUQ scores. Both scales scored out of 100. Average SUS score (68/100) shown in black.

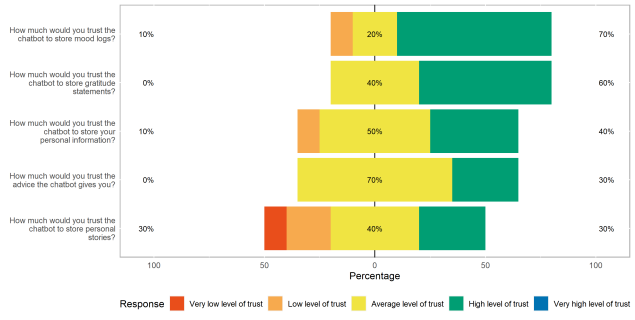

Figure 4: Participant responses to trust questions on ChatPal (n=10). Percentages on left represent very low/ low level of trust (red/ orange). Percentages in the middle indicate average level of trust (yellow). Percentages on right indicate high level of trust (green).

the inclusion of representative users, as study participants were clients of a mental health service with mild-moderate mental health problems and the general population. ChatPal is targeted at both groups. The average SUS score of 64.3±16.8 and CUQ score of 65±6.7 suggest that improvements can be made, and the usability issues found support this. The critical and serious usability issues will be remedied as these stopped participants from using the chatbot at all. The rest of the usability issues were either aesthetic or added functionality. Chatbot personality is an important design consideration which can be challenging. Previous work [26] found that personality traits can still be simulated with a text-based chatbot. To give our chatbot more personality, emojis were used throughout to simulate typical messaging, however it is important that the meaning is clear. For example, emojis were used as buttons ('quick replies'), but this was not clear to participants and should be changed to a 'call-to-action' button. Overall, task 2 had the most usability issues (n=6), but this was mainly a result of software bugs which will be fixed. Grammar errors throughout chatbot conversations should be fixed. The general design recommendations based on this study are as follows:

(1) Slow down the speed of the text to make it more easily digestible, as some of the content appeared too quickly making it difficult for users to read.
(2) The onboarding experience could be improved by having an explainer video which points out the chatbot features and where to find them.
(3) Break down paragraphs of text into smaller chunks to allow for easier content processing. Better use of the user interface layout would also assist with this by displaying timestamps at the end of each dialogue rather than every message.
(4) If asking users to 'rate' the usefulness of the conversation, the rating should appear at the end of each dialogue, before asking the user what they would like to do next. In this case, smiley faces may be a more conventional option like ☺ and ☹ instead of 👍 (good) and 👎 (bad).
(5) Ensure text sizes are consistent between versions of Android, iOS, and the web.

Trust is an important aspect of usability. Overall, participants largely trusted the Chatbot with most responses ranging from average to high trust. Participants trusted the chatbot the most for storing mood logs and gratitude statements and giving advice but were less trusting when it came to the chatbot storing personal data. This emphasises the importance of upholding data privacy, secure data storage and being transparent to the user, as these issues are directly linked to the trustworthiness of digital mental health technologies [16].

Previous work exploring the usability of mental health chatbots [5] suggested that chatbot responses should be variable, to avoid the user receiving the same response for every interaction. This was also noted by participants in the present study. Cameron [5] suggested the chatbot should use emojis and GIFs, which is what we opted for in ChatPal, along with other multimedia (videos). This helps to make the user interactions more interesting, rather than having solely text-based conversations.

A recent review [1] looked at metrics used in studies to evaluate health care chatbots. The authors found that usability was the most frequently evaluated, but most studies used only one question to assess usability, and most did not use SUS. However, other metrics such as CUQ [12] may be more accurate and appropriate for this use case. Future work should seek to build on the methodology utilised in this study, to develop a rigorous usability protocol that can be used to assess digital health tools.

4.1 Limitations

The usability test on the ChatPal chatbot took place during the COVID-19 pandemic so we could not conduct the usability tests in person as initially planned. The interviews were instead completed online using conferencing calls. While this was satisfactory, we do feel that more in depth conversations and reflections could have been made if the study was conducted in person. Additionally, as only a small number of individuals (n=10) were recruited for the study, we were unable to measure demographic differences

in usability and trust across participants. Future work will involve testing the chatbot with more participants across a range of demographics.

5 CONCLUSIONS

The paper reports the findings of a usability study on a prototype of the ChatPal chatbot, designed to support the mental wellbeing of individuals in rural areas. Overall, the participants scored the chatbot below average on the usability scales (SUS and CUQ) and trust in the chatbot was reasonable. The critical issues highlighted were a result of some software bugs which completely stopped participants using the chatbot and forced them to restart the application. The rest of the issues were mainly to do with usability and aesthetic issues, including the speed at which content was delivered, the amount of text delivered at one time and the use of appropriate emojis for rating the conversation. The authors hope that the recommendations provided in this study will be useful for other researchers developing digital health chatbots.

ACKNOWLEDGMENTS

The ChatPal consortium acknowledges the support provided by the Interreg VB Northern Periphery & Arctic Programme under the grant for Conversational Interfaces Supporting Mental Health and Wellbeing of People in Sparsely Populated Areas (ChatPal) project number 345. The authors would like to thank all the clients, participants, project members, supporters, and researchers at Ulster University, University of Eastern Finland, Norrbotten Association of Local Authorities, Region Norrbotten, Luleå University of Technology, NHS Western Isles, Action Mental Health, Munster Technological University, and Health Innovation Hub Ireland, for participating in this research.

REFERENCES

[1] Alaa A. Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M. Bewick, and Mowafa Househ. 2021. Perceptions and Opinions of Patients about Mental Health Chatbots: Scoping Review. *J Med Internet Res* 23, 1 (jan 2021), e17828. https://doi.org/10.2196/17828

[2] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2 (2020), 100006. https://doi.org/10.1016/j.mlwa.2020.100006

[3] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank van der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing* 26, 1 (2022), 95–119. https://doi.org/10.1007/s00779-021-01582-9

[4] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale | john Brooke | Taylor & Fran. https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke.

[5] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2019. Assessing the Usability of a Chatbot for Mental Health Care. In *Internet Science*, Svetlana S Bodrunova, Olessia Koltsova, Asbjørn Følstad, Harry Halpin, Polina Kolozaridi, Leonid Yuldashev, Anna Smoliarova, and Heiko Niedermayer (Eds.). Springer International Publishing, Cham, 121–132.

[6] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (2017), e19. https://doi.org/10.2196/mental.7785

[7] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *Interactions* 24, 4 (jun 2017), 38–42. https://doi.org/10.1145/3085558

[8] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Mental Health* 5, 4 (2018), e64. https://doi.org/10.2196/mental.9782

[9] Paolo Fusar-Poli, Gonzalo Salazar de Pablo, Andrea De Micheli, Dorien H. Nieman, Christoph U. Correll, Lars Vedel Kessing, Andrea Pfennig, Andreas Bechdolf, Stefan Borgwardt, Celso Arango, and Therese van Amelsvoort. 2020. What is good mental health? A scoping review. *European Neuropsychopharmacology* 31 (2020), 33–46. https://doi.org/10.1016/j.euroneuro.2019.12.105

[10] Andrea K. Graham, Ruth Striegel Weissman, and David C. Mohr. 2021. Resolving Key Barriers to Advancing Mental Health Equity in Rural Communities Using Digital Mental Health Interventions. *JAMA Health Forum* 2, 6 (2021), e211149. https://doi.org/10.1001/jamahealthforum.2021.1149

[11] Nidhin Harilal, Rushil Shah, Saumitra Sharma, and Vedanta Bhutani. 2020. CARO: An Empathetic Health Conversational Chatbot for People with Major Depression. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (Hyderabad, India) *(CoDS COMAD 2020)*. Association for Computing Machinery, New York, NY, USA, 349–350. https://doi.org/10.1145/3371158.3371220

[12] 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 2019 - Proceedings of the 31st European Conference on Cognitive Ergonomics: "Design for Cognition"*. Association for Computing Machinery, Inc, New York, NY, USA, 207–214. https://doi.org/10.1145/3335082.3335094

[13] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth* 6, 11 (nov 2018). https://doi.org/10.2196/12106

[14] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. *"I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376175

[15] James R. Lewis and Jeff Sauro. 2018. Item Benchmarks for the System Usability Scale. *J. Usability Studies* 13, 3 (may 2018), 158–167.

[16] Nicole Martinez-Martin. 2020. Chapter Three - Trusting the bot: Addressing the ethical challenges of consumer digital mental health therapy. In *Ethical Dimensions of Commercial and DIY Neurotechnologies*, Imre Bárd and Elisabeth Hildt (Eds.). Developments in Neuroethics and Bioethics, Vol. 3. Academic Press, 63–91. https://doi.org/10.1016/bs.dnb.2020.03.003

[17] Michael McTear. 2020. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots.* Vol. 13. Morgan & Claypool Publishers. 1–251 pages. https://doi.org/10.2200/S01060ED1V01Y202010HLT048

[18] Jakob Nielsen. 1994. Severity Ratings for Usability Problems: Article by Jakob Nielsen. https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/.

[19] Nielson Norman Group. 2012. Thinking Aloud: The #1 Usability Tool. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

[20] Nielson Norman Group. 2012. Usability 101: Introduction to Usability. https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[21] OECD/ EU. 2018. *Health at a Glance: Europe 2018: State of Health in the EU Cycle.* Technical Report. OECD Publishing, Paris. https://doi.org/10.1787/health_glance_eur-2018-en

[22] Claudia Oppenauer, Juliane Burghardt, Elmar Kaiser, Friedrich Riffer, and Manuel Sprung. 2021. Psychological Distress During the COVID-19 Pandemic in Patients With Mental or Physical Diseases. *Frontiers in Psychology* 12 (aug 2021), 703488. https://doi.org/10.3389/FPSYG.2021.703488

[23] PhoneGap.com. 2022. Phonegap.com. http://phonegap.com/products/.

[24] C Potts, E Ennis, R B Bond, M D Maurice, M F McTear, K Boyd, T Broderick, M Malcolm, L Kuosmanen, H Nieminen, A K Vartiainen, C Kostenius, B Cahill, A Vakaloudis, G McConvey, and S O'Neill. 2021. Chatbots to Support Mental Wellbeing of People Living in Rural Areas: Can User Groups Contribute to Co-design? *Journal of Technology in Behavioral Science 2021* (sep 2021), 1–14. https://doi.org/10.1007/S41347-021-00222-6

[25] Rasa.com. 2022. Open source conversational AI | Rasa. https://rasa.com/.

[26] Elayne Ruane, Sinead Farrell, and Anthony Ventresque. 2021. User Perception of Text-Based Chatbot Personality. In *Følstad A. et al. (eds) Chatbot Research and Design. CONVERSATIONS 2020. Lecture Notes in Computer Science*, Vol. 12604. Springer, Cham, 32–47. https://doi.org/10.1007/978-3-030-68288-0_3

[27] Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D Mulvenna, Siobhan O'Neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, Gavin McConvey, Robin Turkington, David Hanna, Heidi Nieminen, Anna-Kaisa Vartiainen, Alison Robertson, and Michael F McTear. 2021. Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts. *ACM Transactions on Computing for Healthcare* 2, 3 (2021), 1–16.

[28] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64, 7 (2019), 456–464.

[29] J. Wainer and J. Chesters. 2000. Rural mental health: neither romanticism nor despair. *The Australian journal of rural health* 8, 3 (2000), 141–147. https://doi.org/10.1046/j.1440-1584.2000.00304.x

[30] World Health Organisation. 2004. *Promoting mental health: concepts, emerging evidence, practice (Summary Report).* Technical Report. World Health Organization, Geneva. https://apps.who.int/iris/bitstream/handle/10665/42940/9241591595.pdf

[31] World Health Organization. 2007. Mental health: strengthening mental health promotion. World Health Organization. http://mindyourmindproject.org/wp-content/uploads/2014/11/WHO-Statement-on-Mental-Health-Promotion.pdf

[32] Jiaqi Xiong, Orly Lipsitz, Flora Nasri, Leanna M.W. Lui, Hartej Gill, Lee Phan, David Chen-Li, Michelle Iacobucci, Roger Ho, Amna Majeed, and Roger S. McIntyre. 2020. Impact of COVID-19 pandemic on mental health in the general population: A systematic review. *Journal of Affective Disorders* 277 (dec 2020), 55–64. https://doi.org/10.1016/J.JAD.2020.08.001