# Assessing the Usability of a Chatbot for Mental Health Care

Gillian Cameron[1,2(✉)], David Cameron[1], Gavin Megaw[1], Raymond Bond[2], Maurice Mulvenna[2], Siobhan O'Neill[3], Cherie Armour[3], and Michael McTear[2]

[1] Inspire Workplaces, 10-20 Lombard Street, Belfast BT1 1RD, UK
{g.cameron,d.cameron,g.megaw}@inspirewellbeing.org
[2] School of Computing, Ulster University, Newtownabbey BT37 0QB, UK
{rb.bond,md.mulvenna,mf.mctear}@ulster.ac.uk
[3] School of Psychology, Ulster University Coleraine Campus, Coleraine BT52 1SA, UK
{sm.oneil,c.armour1}@ulster.ac.uk

**Abstract.** The aim of this paper is to assess the usability of a chatbot for mental health care within a social enterprise. Chatbots are becoming more prevalent in our daily lives, as we can now use them to book flights, manage savings, and check the weather. Chatbots are increasingly being used in mental health care, with the emergence of "virtual therapists". In this study, the usability of a chatbot named iHelpr has been assessed. iHelpr has been developed to provide guided self-assessment, and tips for the following areas: stress, anxiety, depression, sleep, and self esteem. This study used a questionnaire developed by Chatbottest, and the System Usability Scale to assess the usability of iHelpr. The participants in this study enjoyed interacting with the chatbot, and found it easy to use. However, the study highlighted areas that need major improvements, such as Error Management and Intelligence. A list of recommendations has been developed to improve the usability of the iHelpr chatbot.

**Keywords:** Chatbots · Chatbot usability · Mental healthcare · User experience · Human Computer Interaction · Microsoft Bot Framework · System Usability Scale

## 1 Introduction

Chatbots have been defined by Shevat [1] as a new kind of user interface, that can be used for many purposes, such as to book flights [2], purchase goods, and manage savings [3]. Chatbots are becoming increasingly prevalent in society, and it has been predicted that users may soon prefer to engage with chatbots, to complete tasks traditionally done on a web page or mobile application [4].

Voice based chatbots are called upon within mobile devices, computers, and smart speakers such as Amazon Alexa, and Google Home. Text based chatbots can be accessed through many channels, such as Messenger, Kik, Slack and


S. S. Bodrunova et al. (Eds.): INSCI 2018 Workshops, LNCS 11551, pp. 121–132, 2019.
https://doi.org/10.1007/978-3-030-17705-8_11

Telegram, or in a web or mobile application. The user can converse with the chatbot using text or quick replies (buttons).

This paper describes a text based chatbot that has been developed, the iHelpr Chatbot. Usability questionnaires are discussed, and adapted to create a usability test to assess the iHelpr chatbot. This paper is a continuation of previous work completed in the area of chatbots for mental health care [5–7].

### 1.1 Background

The Farmer and Stevenson report sheds light on a significant mental health challenge that the UK faces at work [8]. This report finds that around 300,000 people with a long term mental health problem lose their jobs each year, and around 15% of people that are currently in work have symptoms of a mental health condition. Investors in People produced a report that listed the top five sectors with the most stressed employees, and charities were the third highest [9].

Chatbots are beginning to appear in the area of mental health care. People living in rural communities, or shift workers, may have problems accessing mental health care appointments, and chatbots could be used as a potential solution to this [10]. There is potential to engage students, as Bhakta, Savin-Baden, and Tombs [11] found that students perceived talking to a chatbot as "safe". Chatbots have already been used to support students during periods of exam stress [12]. Woebot [13], a chatbot therapist, has made headlines recently, and receives two million messages per week. A randomised control trial held at Stanford University found students who used Woebot, had significantly reduced symptoms of depression within 2 weeks.

Tess, developed at the company x2-AI by clinical psychologists, delivers support using Cognitive Behaviour Therapy (CBT), Solution-focused Brief Therapy (SFBT), and mindfulness [14].

## 2 Aims

The aims of this study are:

- to assess the usability of a chatbot for mental health care in a workplace setting.
- to develop recommendations to improve the usability of a chatbot for mental health care.

## 3 Inspire Support Hub and iHelpr

In previous works [5–7], Inspire Workplaces developed a chatbot, iHelpr, in partnership with Ulster University through a Knowledge Transfer Partnership. Inspire Workplaces provide programmes and wellbeing solutions for private and

public sector organisations, and educational institutions across the island of Ireland. Inspire Workplaces is a social enterprise, that sits within a wider charity group, called Inspire. To broaden their service offering, and access to their services, Inspire Workplaces has recently developed a digital intervention to complement their existing face-to-face counselling services. The Inspire Support Hub has been developed, which is a website containing self-help tools and resources. The iHelpr chatbot is embedded within the Hub, to guide the user around the self-help tools and resources and provide tailored self-help recommendations.

### 3.1 iHelpr

The iHelpr chatbot provides guided self-assessment on the following topics: stress, anxiety, depression, sleep, and self esteem. iHelpr initially allows the user to complete a self-assessment instrument based on the option they have chosen. Figure 1 shows a screenshot of the iHelpr chatbot issuing questions based on the Perceived Stress Scale to assess perceived stress levels. Tailored advice with evidence-based recommendations are then presented to the user, based on the results of the self-assessment survey [15]. The recommendations include links to other support literature available on the Inspire Support Hub website, and recommended e-learning programmes. If there is escalated risk depending on higher scores, the user is given helpline numbers and if necessary, emergency contact information. The iHelpr chatbot was developed using the Microsoft Bot Framework[1], with NodeJS. It is connected to a MySQL database that holds

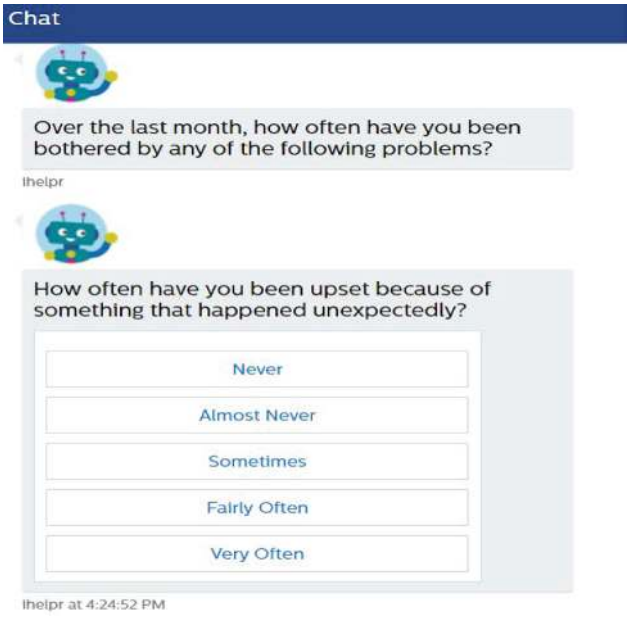


**Fig. 1.** iHelpr

[1] https://dev.botframework.com/.

coping strategies and questionnaire scores. Microsoft's Language Understanding Intelligent Service[2] (LUIS) was incorporated to recognise the utterances made by users and to match them to the correct intent.

**Conversation Design.** The conversation flow has been designed with a clinical psychologist. Conversation scripts were developed, and refined on numerous iterations with the psychologists to ensure they were fit for purpose. The conversation was then inserted into the prototyping tool Botsociety[3], to visualise the conversation flow in a conversational interface. The user can interact with the chatbot using free text in some areas, but mostly through quick replies. Users cannot send GIFs or Emojis to the chatbot, however GIFS are used within the conversation. Randomised GIFs are shown to the user on greeting, as shown in Fig. 2, and on leaving the conversation.

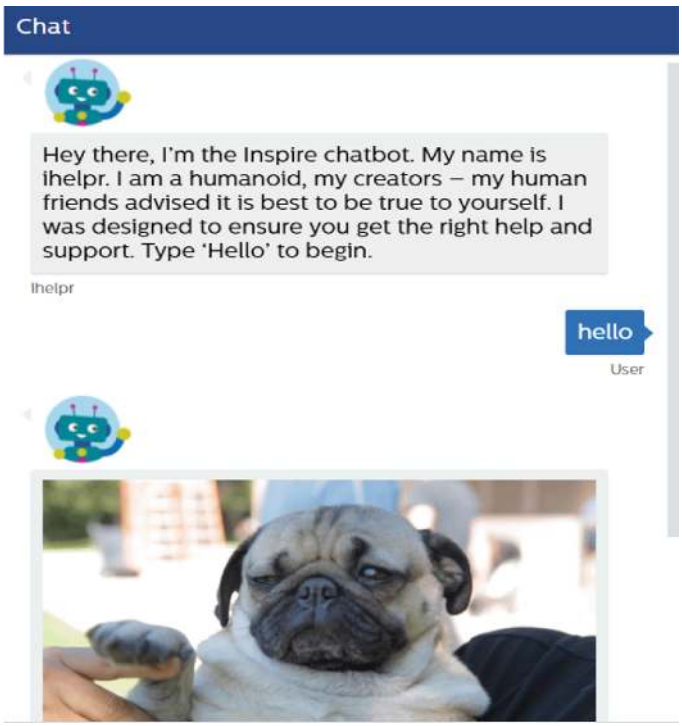


**Fig. 2.** GIFs used within iHelpr

**Onboarding.** The user is prompted to introduce themselves to the chatbot, with the bot asking "What is your name?". The chatbot then onboards the user, explaining the scope and provides a helpline number. The onboarding message the user receives is - "Welcome Gillian! I'm iHelpr! The areas I can help you with are Stress, Anxiety, Self Esteem, Sleep and Depression. Type 'Menu' at any time to view options. If you are in immediate need, please call our helpline on 0800 389 5362. Type Continue to move on with our chat:)".

[2] https://www.luis.ai.
[3] https://botsociety.io/.

### 3.2 Inspire Support Hub

The iHelpr chatbot is embedded within an online self-help portal called the Inspire Support Hub. The features of the Inspire Support Hub include;

- **five ways to wellbeing database:** a searchable database, where users can find resources and groups in their area, based on the Five Ways to Wellbeing which are Be Active, Keep Learning, Take Notice, Give, and Connect with people [16].
- **online self-help library:** reading materials on common mental health conditions including stress, anxiety, depression.
- **e-library of bibliotherapy books:** books on a range of self-help topics.
- **elearning programmes:** on stress, anxiety and depression, sleep, self-esteem and alcohol.
- **thought diary function:** users can track their moods, thoughts and input journal entries.

## 4 Related Work

### 4.1 Usability Questionnaires

The term usability is part of a broader term - "user experience". Usability refers to how easy it is to access or use a system. Questionnaires have been observed as the most frequently used tools for usability evaluation. The section below describes three of the most frequently used questionnaires:

**The USE Questionnaire** developed by Lund, measures Usability, Satisfaction and Ease of use [17]. The questionnaire consists of 30 questions and utilises a seven point Likert rating scale, ranging from −3 totally disagree to +3 totally agree.

**Software Usability Measurement Inventory** (SUMI) consists of 50 questions, and provides an objective way of assessing user satisfaction with a piece of software [18]. The responses utilise a three point scale - agree, undecided, disagree.

**The System Usability Scale** (SUS) was developed by Brooke in 1996 [19], and is described as a "quick and dirty" usability scale. It is widely used and allows the researcher to quickly and easily assess the usability of a system. SUS contains 10 questions and participants respond by selecting one of five points that range from strongly disagree to strongly agree. A SUS score ranges between 0–100 and a score above 68 is classed as above average. SUS is very flexible, and can be applied to a wide variety of interfaces, including websites and voice response systems. More recently, it has been adapted to evaluate the usability of chatbot platforms [20].

### 4.2 Chatbot Usability Testing

Kocaballi, Laranjo and Coiera [21] compared the following questionnaires for measuring user experience in conversational interfaces:

1. AttrakDiff questionnaire; which measures how attractive a product is based on it's hedonic and pragmatic qualities.
2. The Subjective Assessment of Speech System Interfaces (SASSI) questionnaire; which can be used to evaluate speech input quality in speech interfaces.
3. The Speech User Interface Service Quality (SUISQ); which assesses the quality of speech interfaces.
4. MOS-X; which contains 15 items to assess how natural synthetic voices are.
5. The System Usability Scale (SUS); a likert scale questionnaire to assess the ease of use of a system.
6. The Paradigm for Dialogue Evaluation System (PARADISE); which is a framework for assessing user satisfaction.

In their study, it was found that a blend of questionnaires was needed to measure chatbot usability.

Chatbottest has developed a collaborative guide of questions, that fall under 7 different categories to test the specific functionality of chatbots [22]. The 7 categories are: Answering, Error management, Intelligence, Navigation, Onboarding, Personality and Understanding. Chatbottest has built a Chrome extension to test chatbots, that uses the collaborative guide of questions and returns an overall percentage. This percentage is based on how well the chatbot scores on the seven different categories. Furthermore, a report is generated with tips on the areas of your chatbot that need improvement.

## 5 Methods

### 5.1 Participants

The participants comprised of 7 employees from a mental health social enterprise. 5 employees were female, 2 were male, 4 were aged between 25 and 34, and 3 between 35 and 44. All employees who participated were in full time employment.

### 5.2 Procedure

Information sheets about the study were given to the participants, consent forms were signed and any questions from the participants were answered. Demographic information was collected prior to beginning the usability test. Usability studies lasted no longer than 40 min per session. Participants were informed their data would be anonymised. All data was anonymised, with participants given IDs such as participant 1, participant 2. The method used to evaluate usability of the chatbot is the open source questions developed by Chatbottest. Many of the Chatbottest questions start with tasks to perform using the chatbot, followed up by a question. The participants were asked to interact with the chatbot using

a PC. In the Chatbottest questionnaire there are 29 questions, and many of the questions have yes/no answers. A screenshot of the start of the Chatbottest questionnaire is shown in Fig. 3.

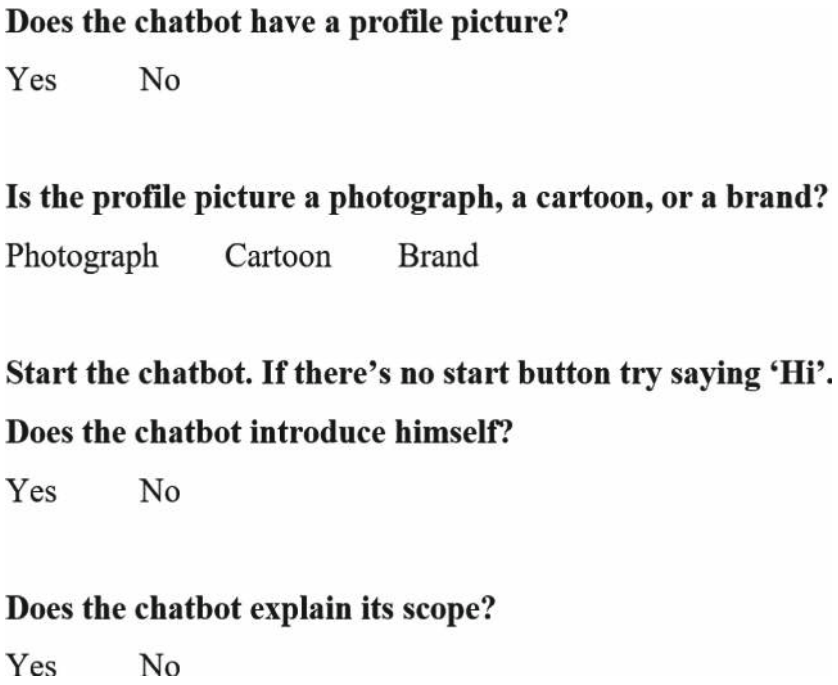


**Fig. 3.** Chatbottest questions

Once the user completed the Chatbottest questionnaire, they were asked to fill out a SUS questionnaire. As the SUS questions are easily adapted for use with different types of systems, and contains only 10 questions, it was chosen for this study. The SUS questions used in this study are listed below:

1. I think that I would like to use this Chatbot frequently.
2. I found the Chatbot unnecessarily complex.
3. I thought the Chatbot was easy to use.
4. I think that I would need the support of a technical person to be able to use this Chatbot.
5. I found the various functions in this Chatbot were well integrated.
6. I thought there was too much inconsistency in this Chatbot.
7. I would imagine that most people would learn to use this Chatbot very quickly.
8. I found the Chatbot very cumbersome to use.
9. I felt very confident using the Chatbot.
10. I needed to learn a lot of things before I could get going with this Chatbot.

The usability tests were conducted on one-to-one basis, with a researcher observing the participant interact with the chatbot, whilst the participant filled in the questionnaire. Verbal comments throughout the test were recorded. Quantitative data were collected from the two questionnaires, and qualitative data from comments made by the participants while using the chatbot.

# 6 Results

## 6.1 SUS Scores

SUS Scores were calculated following the guidelines given by Brooke [19]: For items 1, 3, 5, 7, and 9 subtract 1 from the score given. For items 2, 4, 6, 8, and 10, subtract the score given from 5. Finally, multiply the sum of the scores by 2.5 to get the final score. The average SUS score for the iHelpr Chatbot was 88.2, which is above the average industry score of 68. Two participants gave the iHelpr chatbot 100, and only one participant gave a score that was under the average of 68. Each participant's SUS score is displayed in Fig. 4.

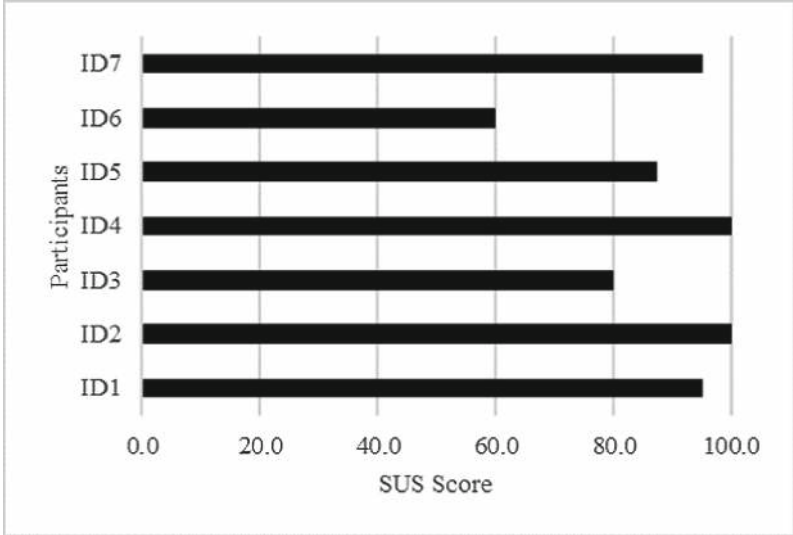


**Fig. 4.** SUS scores

## 6.2 Chatbottest Questionnaire

An overall percentage rating is calculated using the Chatbottest chrome extension. Scores out of 100 for each category, Onboarding, Personality, Chatbot Answering, Chatbot Understanding, Navigation, Error Management and Intelligence are also calculated depending on the answers the user gives. The average percentage for the iHelpr Chatbot was 55.6%. The lowest result was 43% and the highest was 74%.

The highest performing categories were Personality, and Onboarding which both scored 100% across all 7 usability tests. Chatbot Answering scored an average of 89% over the 7 tests, however some categories scored very low, with Chatbot Understanding scoring 24% on average, and Error Management scoring 14%. A chart plotting the average scores for each category is shown in Fig. 5. Chatbottest also provide a report detailing tips on how to improve the chatbot based on the participants responses.

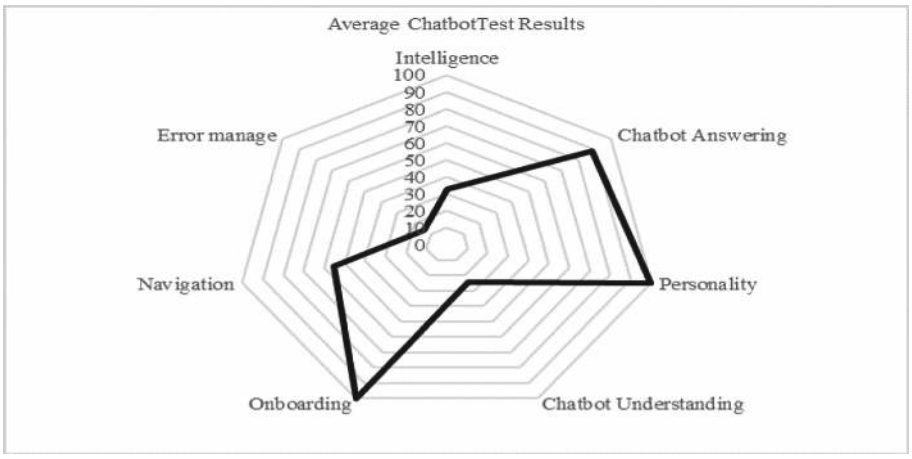


**Fig. 5.** iHelpr Chatbot performance on Chatbottest

4 out of 7 participants took less than 4 steps to get something valuable out of the chatbot. All participants found that elements such as images, emojis, and Graphic Interchange Format images (GIFs) were not understood by the chatbot, and neither was typing in different languages. 5 out of 7 participants found that the chatbot could maintain a conversation when asked generic questions, such as "How are you" or "Can you tell me a joke?". However, the variability in the chatbot responses was found to be minimal, with 5 out of 7 participants answering "No" to "Say something not really nice to the chatbot. You can start with something like 'idiot' and then go further from there. Does it have answers for your different bad words?". Furthermore, 5 out of 7 participants did not receive a variation of answers when saying "nice" things to the chatbot. All participants found that when they typed a word incorrectly, such as "anious" instead of "anxious" the chatbot did not understand. The participants found there was good use of high quality elements such as GIFs throughout the conversation, as well as rich media elements such as quick reply buttons. The voice and tone of the chatbot were perceived to be consistent, and participants thought the personality of the chatbot was the correct fit for the intended audience.

## 7 Discussion

The results of this study indicate the chatbot performs well on some areas, such as Onboarding and Personality. The results highlight areas for much needed improvement, in how the iHelpr Chatbot handles errors, and how user requests are understood successfully. The SUS scores are promising, and provide a basis to improve on, and compare with future studies. In Coperich, Cudney and Nembhard's study, SUS was utilised to assess the usability of chatbot development platforms, Watson and Pandorabots [20]. Watson scored an average SUS score of 81.875 and Pandorabot scored 88.75.

During the usability tests, comments were made on the personality of the chatbot, with participant 1 stating iHelpr was "Upbeat" and participant 3 stating the chatbot was "friendly". Participant 1 and 4 both expressed that they "liked" the chatbot. Participant 7 said the chatbot was "amazing". Personality can be a crucial factor in determining whether the user wants to utilise the chatbot again [23]. The chatbottest questionnaire contains a questions around the voice and tone of the chatbot - "Can you identify a specific voice and tone in the chatbot that is consistent throughout the conversation? Do you think the voice and tone fits with the purpose of the chatbot?" Two participants found these questions confusing, as they thought it was more applicable to a voice based chatbot, such as an Amazon Alexa rather than a text based chatbot. For future use, this question could possibly be rephrased depending on the chatbot being evaluated.

When asked the question "Ask the chatbot about common daily stuff. Things like: How are you? - Where are you from? Tell me a joke", participants all asked the chatbot to tell them a joke. The chatbot responded appropriately with randomised jokes, which all participants laughed at. The chatbot sends GIFs to the user at different points in the conversation, and participant 4 stated that these "cheered them up", and participant 7 particularly liked the "animal GIFs". Participants were unable to send emojis or GIFs back to the chatbot, which will need to be addressed to improve the conversational experience. Participant 2 and 4 found the usability test "fun" and liked interacting with the chatbot. When asked to type in another language to the chatbot, participants found this functionality was not supported, and participant 1 stated this would be a very useful feature to implement.

One of the questions in the Chatbottest questionnaire is - "Say something not really nice to the chatbot. Does it have answers for your different bad words?". The chatbot responded with, "That's not very nice - I am only trying to help." Participant 3 stated this could be phrased better, with a response such as "I'm sorry you feel that way". Participant 2 stated that the use of the helpline number in different points of the conversation was beneficial, as it makes users aware they can phone a mental health care professional at any time. Error management in iHelpr requires immediate improvement, as misspelled words are not detected. This could be improved by utilising spell checking APIs, such as Bing Spell Check[4] which can be integrated with LUIS. This allows the mistyped word to be corrected, before being sent to LUIS to predict the intent. Variability in the chatbot responses would ensure the user does not receive the same error message on each occasion.

A limitation of this study is that it was not known if participants were actively experiencing common mental health issues. Many participants stated they would only use the chatbot if they felt like they needed it, therefore a further study with participants who are actively experiencing a mental health issue would need to be undertaken. Another limitation was the small sample size, as hypothesis formed from this study will need further confirmation through future studies, such as monitoring real-world use of the iHelpr chatbot.

[4] https://azure.microsoft.com/en-gb/services/cognitive-services/spell-check/.

### 7.1 Final Recommendations

Drawing on the results and discussion, recommendations to improve the iHelpr chatbot have been derived. The utterances typed during the usability study should be compiled into a dataset to train the chatbot. More variability should be added to the chatbot responses, so that the user does not receive the exact same responses on each interaction. A more robust error management strategy needs to be developed to counteract the errors found during the usability study. A Bing Spell Check API should be integrated with LUIS to correct mistyped words. Functionality to allow the user to interact with the chatbot using Emojis, GIFs, and other elements needs to be developed. Localization should be supported to allow the user to interact with the chatbot in multiple languages. Another usability study should be completed to ensure the issues found are rectified. Furthermore, a usability study with participants who are experiencing a mental health problem should be undertaken.

## 8 Conclusion

In conclusion, the participants found the iHelpr Chatbot to be enjoyable and easy to use, and stated there is a consistent personality throughout the conversation, and the chatbot performs well at onboarding. Error Management and Intelligence are areas that require urgent attention, as they performed poorly in the questionnaire. Using the chatbottest questionnaire and SUS together was found to be a good combination, as many participants enjoyed completing the usability study. This study was completed with participants in a social enterprise. A further study would need to be undertaken with a larger selection of participants who are experiencing a mental health issue, to assess if the iHelpr Chatbot is useful in this context.

**Acknowledgements.** This study has been supported by UK Knowledge Transfer Partnership under KTP grant ID 1022267.

## References

1. Shevat, A.: Designing Bots: Creating Conversational Experiences. O'Reilly Media Inc., Newton (2017)
2. Kayak. https://www.kayak.com/messenger. Accessed 2 Aug 2018
3. Plum. https://withplum.com/. Accessed 2 Aug 2018
4. Følstad, A., Brandtzæg, P.B.: Chatbots and the new world of HCI. Interactions **24**(4), 39–42 (2017)
5. Cameron, G., et al.: Towards a chatbot for digital counselling. In: Proceedings of the 31st British Computer Society Human Computer Interaction Conference, p. 24. BCS Learning & Development Ltd., Sunderland (2017). http://dx.doi.org/10.14236/ewic/HCI2017.24
6. Cameron, G., et al.: Best practices for designing chatbots in mental healthcare - a case study on iHelpr. In: Proceedings of the 32nd Human Computer Interaction Conference (2018). http://dx.doi.org/10.14236/ewic/HCI2018.129

7. Cameron, G., et al.: Back to the future: lessons from knowledge engineering methodologies for chatbot design and development. In: Proceedings of the 32nd Human Computer Interaction Conference (2018). http://dx.doi.org/10.14236/ewic/HCI2018.153
8. Stevenson, D., Farmer, P.: Thriving at work: the Stevenson/Farmer review of mental health and employers (2017)
9. Investors in People: Managing Mental Health in the Workplace 2018 (2018)
10. Miner, A., et al.: Conversational agents and mental health: theory-informed assessment of language and affect. In: Proceedings of the Fourth International Conference on Human Agent Interaction, pp. 123–130. ACM (2016). https://doi.org/10.1145/2974804.2974820
11. Bhakta, R., Savin-Baden, M., Tombs, G.: Sharing secrets with robots? In: EdMedia: World Conference on Educational Media and Technology, pp. 2295–2301. Association for the Advancement of Computing in Education (AACE) (2014)
12. Kavakli, M., Li, M., Rudra, T.: Towards the development of a virtual counselor to tackle students' exam stress. J. Integr. Des. Process Sci. **16**(1), 5–26 (2012). https://doi.org/10.3233/jid-2012-0004
13. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: 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). https://doi.org/10.2196/mental.7785
14. X2ai.com. http://x2ai.com/. Accessed 2 Aug 2018
15. Cohen, S., Kamarck, T., Mermelstein, R.: Perceived stress scale. In: Measuring Stress: A Guide for Health and Social Scientists. Oxford University Press, Oxford (1994)
16. New Economics Foundation. https://neweconomics.org/2008/10/five-ways-to-wellbeing-the-evidence/. Accessed 2 Aug 2018
17. Lund, A.M.: Measuring usability with the USE questionnaire. STC Usability SIG Newsl. **8**(2), 3–6 (2001)
18. Kirakowski, J., Corbett, M.: SUMI: the software usability measurement inventory. Br. J. Educ. Technol. **24**(3), 210–212 (1993)
19. Brooke, J.: SUS: a "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor and Francis, London (1996)
20. Coperich, K., Cudney, E., Nembhard, H.: Continuous improvement study of chatbot technologies using a human factors methodology. In: Proceedings of the 2017 Industrial and Systems Engineering Conference (2017)
21. Kocaballi, A.B., Laranjo, L., Coiera, E.: Measuring user experience in conversational interfaces: a comparison of six questionnaires. In: Proceedings of the 32nd Human Computer Interaction Conference (2018). http://dx.doi.org/10.14236/ewic/HCI2018.21
22. Chatbottest. http://chatbottest.com/. Accessed 3 Aug 2018
23. Callejas, Z., López-Cózar, R., Ábalos, N., Griol, D.: Affective conversational agents: the role of personality and emotion in spoken interactions. In: Conversational Agents and Natural Language Interaction: Techniques and Effective Practices, pp. 203–222. IGI Global, Hershey (2011). https://doi.org/10.4018/978-1-60960-617-6.ch009