Perceptions of Chatbots in Therapy

Samuel Bell
University of Cambridge
Cambridge, UK
sjb326@cam.ac.uk

Clara Wood
Ieso Digital Health
Cambridge, UK
c.wood@iesohealth.com

Advait Sarkar
Microsoft Research
Cambridge, UK
advait@microsoft.com

ABSTRACT

Several studies have investigated the clinical efficacy of remote-, internet- and chatbot-based therapy, but there are other factors, such as enjoyment and smoothness, that are important in a good therapy session. We piloted a comparative study of therapy sessions following the interaction of 10 participants with human therapists versus a chatbot (simulated using a Wizard of Oz protocol), finding evidence to suggest that when compared against a human therapist control, participants find chatbot-provided therapy less useful, less enjoyable, and their conversations less smooth (a key dimension of a positively-regarded therapy session). Our findings suggest that research into chatbots for cognitive behavioural therapy would be more effective when directly addressing these drawbacks.

INTRODUCTION

Mental illness is a leading contributor to the global health burden, with approximately one third of people experiencing poor health in their lifetime [2]. An effective treatment option is Cognitive

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05…$15.00
https://doi.org/10.1145/3290607.3313072

Behavioural Therapy (CBT). However, access to treatment often entails long wait lists, high cost, and logistical difficulties (e.g., time and transport requirements).

To provide scalable treatment, several promising studies have demonstrated clinical efficacy of internet-based CBT, whereby the need for a face-to-face presence is negated [1, 5]. However, limited numbers of specialist therapists still present a bottleneck in service provision [5]. Woebot, a template-based chatbot delivering basic CBT, has demonstrated limited but positive clinical outcomes in students suffering from symptoms of depression [4]. Here, we seek to determine differences in user *perception* of a therapy session if provided by a chatbot.

We conducted a between-subjects study with 10 participants who self-identified with sub-clinical symptoms of stress. Each group engaged in two 30-minute sessions of internet CBT, the first group with a human therapist, and the second with a "chatbot" therapist, which in reality was simulated using the Wizard of Oz technique. Perceptions of each session were evaluated using the Session Evaluation Questionnaire (SEQ) [9], alongside a separate study-specific questionnaire, and followed up with a qualitative interview.

H1 The perceived *sharing ease* reported by participants is affected by whether they were talking with a chatbot or a human therapist.

H2 The perceived *smoothness* reported by participants is affected by whether they were talking with a chatbot or a human therapist.

H3 The perceived *usefulness* reported by participants is affected by whether they were talking with a chatbot or a human therapist.

H4 The perceived *enjoyment* reported by participants is affected by whether they were talking with a chatbot or a human therapist.

Sidebar 1: Hypotheses H1 to H4 are evaluated in this work.

METHOD

This study aims to assess the potential for chatbot therapists as a mechanism to reduce cost, increase access to treatment, and overcome the practitioner bottleneck. While clinical efficacy is critical for such therapists there are other factors that contribute to a good therapy session [3, 9]. In this work, we focus on four metrics of session perception, leaving clinical efficacy for future study. The first, here termed "sharing ease", refers to a participant's perception of being able to confide in their therapist and share personal information, and is measured principally by SEQ.[1] The second, termed "smoothness", refers to how flowing and easy the conversation is, and is also measured by SEQ. The third, termed "usefulness", refers to the participant's own perception of the session's efficacy, measured by response to Likert-scale questionnaire. Finally, "enjoyment" refers to how enjoyable the session is perceived as, also measured by questionnaire. We introduce the latter two measures, on the basis that sessions perceived as useful and enjoyable are likely to result in high adherence. These four metrics correspond to the hypotheses in Sidebar 1.

Experiment design

In order to assess these hypotheses, we pilot a randomised controlled trial with two groups. Group A, the control, chatted with a human therapist through an internet-based CBT chat interface, developed by Ieso Digital Health. Group B chatted with a "chatbot" therapist through the same interface. Both groups were informed as such. However, as no suitably advanced therapy chatbot exists today, we developed a Wizard of Oz setup.

[1] We use the term "sharing ease" over "depth" as in [9] because we also consider responses to a Likert-scale questionnaire.

Both groups participated in two 30-minute sessions of internet-based CBT. While one benefit of internet CBT is the option of remote access, in this study all participants took part in their therapy whilst under supervision in an experimental suite.

In order to simulate the chatbot therapist, a conversation script was devised, following inspiration from conversation scaffolds for peer-support [7]. Such a script enumerates possible responses of a chatbot therapist. Template variables are inserted at key points in the script, such that extracts of participant responses can be substituted in for later repetition. During a chatbot therapy session, the chatbot is 'driven' by a trained psychotherapist and a researcher, co-selecting the most appropriate answer from the script at a given time. See Sidebar 2 for an example conversation.

The freedom of the chatbot 'wizard' is carefully limited. Several catch-all responses can be used when a participant cannot be answered with a response from the script.

Chatbot Do you think that there is a place where we can change your behaviour to see if we can change your feelings?

Participant I could try going to bed at a fixed time each night.

Chatbot That's a great aim, *Jimmy. Going to bed at a fixed time each night* would be a great way to try to reach this goal.

Sidebar 2: An example conversation between the chatbot and a participant (template substitutions *italicised*).

Participant recruitment

We recruited 10 participants via posters around departments and colleges of the University of Cambridge, and posts on social media. Participants were invited to apply if they self-identified with symptoms of stress and had no previous diagnosis, nor received treatment for, a mental health condition. All participants were pre-screened for clinical risk with the Patient Health Questionnaire (PHQ-9) [6]. Participants were then randomly assigned to the two groups, resulting in 5 participants per group.

Ethics

This study was approved by the School of Technology Ethics Committee at the University of Cambridge. All participants gave written informed consent.

Data collection and evaluation

After each therapy session, participants were asked to evaluate their experience with two questionnaires. The first is the SEQ, comprising 22 semantic differential scale questions. A typical question might ask the participant to respond to "The session was: Relaxed—Tense", choosing an integer from 1 to 7 where 1 represents completely relaxed and 7 represents completely tense. Two standard SEQ metrics, "depth", and "smoothness", can be derived from these responses. Full details on this derivation can be found in [9].

The second comprises three statements (see Sidebar 3) to which the participant must respond with a typical 5-point Likert scale of answers, ranging from "Strongly disagree" to "Agree".

(1) "I enjoyed the session today."
(2) "I felt that the session today was useful."
(3) "I felt that I could share openly with my therapist."

Sidebar 3: Agree-disagree statements from the post-session questionnaire.

H1 anticipates a difference in mean responses to statement 3 of the Likert questionnaire, and a difference in SEQ depth. H2 predicts a difference in mean SEQ smoothness. H3 a difference in mean responses to Likert statement 2, when considered interval data. H4 predicts a difference in mean responses to Likert statement 1. Our significance threshold is 0.05.

"It very much felt like a bot; it felt like it was just picking up on keywords I was saying."

"You'd still be able to tell it's a robot."

"I didn't feel that I could be any different to speaking with someone, really."

Sidebar 4: Participant quotes validating our experiment design.

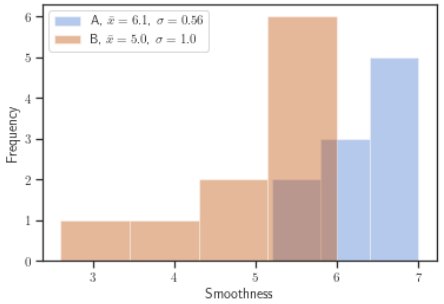

Figure 1: Histogram of SEQ smoothness.

P12: "It was repetition of what I said, not an expansion of what I said"

P1: "I felt standard answers come back … *anybody* could say that"

P2: "It suggested keeping a thought journal, but then it didn't really expand on what it meant."

Sidebar 5: Participant comments regarding chatbot answers.

To add detail, a semi-structured interview was conducted after completion of the questionnaire. Participants were asked to describe their perception of the therapy.

RESULTS

From the informal interviews, we established that all members of Group B believed they were speaking to a chatbot. Without interviewer prompting, all members of Group B confirmed their belief that their therapist was indeed a chatbot. The quotes in Sidebar 4 are examples.

Sharing ease

To evaluate sharing ease, we first consider SEQ depth. Group A report a mean depth of 4.5 ± 0.52, and Group B a mean of 4.2 ± 0.41. This difference is not statistically significant (Welch's t-test, p=0.252).

We consider Likert-scale responses as interval data, mapping to an integer between -2 and $+2$, with "Strongly Disagree" at -2 and "Strongly Agree" at $+2$. These were not normally distributed (according to the Shapiro-Wilk test), so we compare median values and apply the Mann-Whitney U test to assess significance.

Measuring agreement with the statement "I felt that I could share openly with my therapist", Group A report a median score of 1.5, and Group B 1.0. The difference is not statistically significant (p=0.07).

Despite a lack of statistical significance, two participants reported differences in perceived sharing ease in our interviews. P12 (Group B) also reports a lack of empathy, and comments on the lack of shared experience of their chatbot interlocutor:

> "When you tell something to someone, it's better, because they might have gone through something similar … there's no sense that the robot cares or understands or empathises."

Smoothness

Group A report a mean SEQ smoothness of 6.1 ± 0.56. Group B report a significantly lower mean of 5.0 ± 1.0 (Welch's t-test, p=0.011). See Figure 1.

Perhaps due to this decreased smoothness, 3 of 5 Group B participants comment on the difficulties of chat-based CBT, with no Group A participants making such an observation:

> P12: "Text is not always as nice as sitting down to something face-to-face, especially with body language."

We thus find support for H2: participants talking with a human therapist experience smoother conversations.

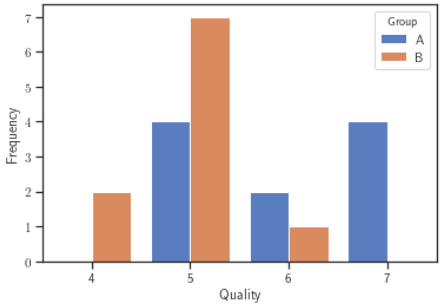

Figure 2: Histogram of 1—7 responses to the statement "The session was: Bad—Good", with 1 representing entirely bad and 7 representing entirely good.

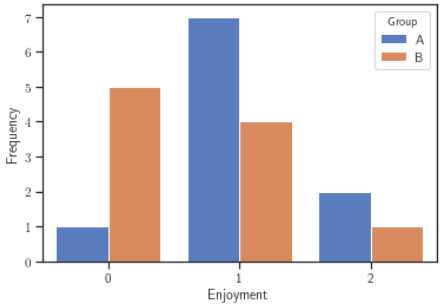

Figure 3: Histogram of Likert-scale agreement with the statement "I enjoyed the session today."

Usefulness

In Likert-scale agreement with the statement "I felt that the session today was useful", both Group A and Group B report an identical median agreement of 1.0.

From interview data, we observe only a minor difference between the two groups. All Group B participants commented at least once that a session had been useful, such as P2:

> "It gave me some good resources …, so in those terms, yes it was useful."

In contrast, only 3 of 5 participants in Group A made remarks about the usefulness of the session.While this data is not strong evidence of a difference in perceived usefulness, we do observe that Group B participants report awareness of chatbot limitations, especially with regard to superficial, standardised answers and the usefulness of suggested goals. Sidebar 5 contains examples.

Enjoyment

In Likert-scale agreement with the statement "I enjoyed the session today", Group A report a median agreement of 1.0. Group B report a median agreement of 0.5. The difference is significant (p=0.05, Mann-Whitney U test). See Figure 3.

Also informative are participant responses the SEQ question "The session was: Bad—Good". Here, Group A report a median of 6.0. Group B report a median of 5.0. The difference is statistically significant (p=0.01, Mann-Whitney U test). See Figure 2.

From both agreement with the statement "I enjoyed the session today" and response to the Good—Bad SEQ scale, we find evidence to support H4.

DISCUSSION

Sidebar 6 summarises our results. We found no evidence in favour of chatbot CBT providing an improved experience, and in some cases find evidence of chatbot CBT resulting in a degraded session perception. There are a few potential reasons for these findings.

First, a strong patient-therapist relationship is built on shared trust, which is typically developed over multiple sessions. Qualitative findings in this work already suggest a lack of empathy, trust and sense of relationship in chatbot CBT.

Second, we posit that Group B were not presented with adequate opportunity for disclosure. With a poor ability to "read between the lines" (i.e., to contextualise and infer meaning beyond the literal conversation), a chatbot may leave its interlocutor feeling that their comment was ignored. This is evidenced by comments such as "It was repetition of what I said, not an expansion of what I said".

Third, the lack of smoothness in conversation is a natural limitation of today's chatbot technology, and the limited degrees of freedom in our script reflect this. This is also likely to be a key reason for the limited usefulness of chatbot responses, particularly with regard to goal suggestions.

H1: sharing ease We found no evidence supporting H1.

H2: smoothness We found evidence supporting H2.

H3: usefulness We found no evidence supporting H3, but qualitative results were suggestive of a potential effect.

H4: enjoyment We found evidence supporting H4.

Sidebar 6: Summary of results.

ACKNOWLEDGEMENTS

We would like to thank all at Ieso Digital Health for their invaluable support, advice, and access to resources.

Future work

As our small cohort cannot reasonably support segmentation analysis, a fruitful area of future investigation would be exploring how different populations (e.g. age groups, or specific diagnoses) respond to chatbot therapy.

We observe that no participants report problems with chatbot recall, though it is difficult for our chatbot to refer to prior sessions. With further sessions this may affect the perception of the therapy.

Finally, we have noted difficulties regarding empathy and sense of shared experience. Without such faculties, no automated therapist will ever come close to building the strength of relationship required for effective therapy, especially given the common desire for genuine social interaction throughout internet CBT [8].

We suggest that future research into chatbot CBT acknowledges and explores these areas of conversational recall, empathy, and the challenge of shared experience, in the hope that we may benefit from scalable, accessible therapy where needed.

REFERENCES

[1] Gerhard Andersson, Pim Cuijpers, Per Carlbring, Heleen Riper, and Erik Hedman. 2014. Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. *World Psychiatry* 13, 3 (2014), 288–295.

[2] Laura Andrade, JJ Caraveo-Anduaga, Patricia Berglund, R Bijl, RC Kessler, Olga Demler, Ellen Walters, C Kylyc, D Offord, TB Ustün, et al. 2000. Cross-national comparisons of the prevalences and correlates of mental disorders. *Bulletin of the World Health Organization* 78 (2000), 413–425.

[3] Orlinsky DE and Howard KI. 1967. The good therapy hour: Experiential correlates of patients; and therapists; evaluations of therapy sessions. *Archives of General Psychiatry* 16, 5 (1967), 621–632. https://doi.org/10.1001/archpsyc.1967.01730230105013 arXiv:/data/journals/psych/12150/archpsyc$_1$6$_5$013.pdf

[4] Kara Kathleen 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 Ment Health* 4, 2 (06 Jun 2017), e19.

[5] Alan E. Kazdin and Stacey L Blase. 2011. Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. *Perspectives on psychological science : a journal of the Association for Psychological Science* 6 1 (2011), 21–37.

[6] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals* 32, 9 (2002), 509–515.

[7] Katie O'leary, Stephen M Schueller, Jacob O Wobbrock, and Wanda Pratt. 2018. Suddenly, we got to become therapists for each other: Designing Peer Support Chats for Mental Health. (2018). https://students.washington.edu/kathlo/SupportiveChat-CHI18-CameraReady.pdf Preprint.

[8] Stefan Rennick-Egglestone, Sarah Knowles, Gill Toms, Penny Bee, Karina Lovell, and Peter Bower. 2016. Health Technologies' In the Wild': Experiences of Engagement with Computerised CBT. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2124–2135.

[9] William B Stiles. 1980. Measurement of the impact of psychotherapy sessions. *Journal of consulting and clinical psychology* 48, 2 (1980), 176.