Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy

Tahir Abbas,[1,2] **Vassilis-Javed Khan,**[1] **Ujwal Gadiraju,**[3] **Panos Markopoulos**[1]
[1]Industrial Design Department, Eindhoven University of Technology, Eindhoven, Netherlands
[2]Software Engineering Dept., Mirpur University of Science & Technology, Mirpur AJK, Pakistan
[3]Web Information Systems, Delft University of Technology, Netherlands
{t.abbas, v.j.khan, p.markopoulos}@tue.nl, u.k.gadiraju@tudelft.nl

Abstract

On-demand emotional support is an expensive and elusive societal need that is exacerbated in difficult times – as witnessed during the COVID-19 pandemic. Prior work in affective crowdsourcing has examined ways to overcome technical challenges for providing on-demand emotional support to end users. This can be achieved by training crowd workers to provide thoughtful and engaging on-demand emotional support. Inspired by recent advances in conversational user interface research, we investigate the efficacy of a conversational user interface for training workers to deliver psychological support to users in need. To this end, we conducted a between-subjects experimental study on Prolific, wherein a group of workers (N=200) received training on motivational interviewing via either a conversational interface or a conventional web interface. Our results indicate that training workers in a conversational interface yields both better worker performance and improves their user experience in on-demand stress management tasks.

Introduction

Coping with stress is crucial for a healthy lifestyle. Prolonged and high levels of stress in humans can affect several physiological and psychological functions (Taelman et al. 2009; Joëls et al. 2006). The recent outbreak of COVID-19 can further affect mental health of people who may fear infection or infecting others, social isolation, sickness and loss of a loved one, among other reasons (Taylor et al. 2020). Recent advances in AI have led to the development of technological interventions for treating stress and anxiety (Shingleton and Palfai 2016). Two potential benefits of such systems for the users are: 1) self-disclosure can be easier as users may have less concern for negative evaluations in the case of a virtual agent (Lucas et al. 2014); 2) therapeutic support can be affordable and more easily available on-demand for a wider range of people in need. Nevertheless, building a fully autonomous virtual therapist for delivering psychotherapeutic solutions is a very challenging endeavor, which requires advancing research in emotional intelligence, affect analysis, computational psychology, and Automatic Speech Recognition (Vogel and Morgan 2009), among other fields.

Crowd-powered solutions, however, can circumvent many of the aforementioned challenges. For example, researchers in *affective crowdsourcing* (Morris 2011) have already proposed methods to leverage crowdsourcing to deliver positive psychological interventions to people who are stressed (Morris and Picard 2012). Another example is Panoply (Morris, Schueller, and Picard 2015), a crowd-powered system that leverages the crowds' wisdom to provide on-demand emotional support to people. Nevertheless, affective crowdsourcing brings about another challenge – that of training workers who have little to no domain knowledge, for delivering on-demand therapeutic support.

Existing training methods in crowdsourcing primarily focus on teaching workers how to solve prevalent microtasks such as complex web search (Doroudi et al. 2016), writing consumer reviews (Dow et al. 2012), information finding (Gadiraju and Dietze 2017), and sentiment analysis tasks (Gadiraju, Fetahu, and Kawase 2015). A notable effort to train workers in the domain of affective crowdsourcing, is a short training method to teach workers only two aspects: expressing empathy and recognizing distortions in input stressors (Morris 2015). How can one train a non-expert worker to deliver positive psychological support? Such training is challenging since it requires a plethora of skills ranging from understanding a person's thoughts and feelings to deciding what actions to undertake based on specific problems.

We investigate how to train workers to solve therapeutic tasks with the help of a conversational interface guided by a chatbot. Chatbots are attracting considerable interest due to their ability to interact with humans in a natural way (Moore et al. 2017). Recently, researchers have investigated the suitability of conversational interfaces for microtask crowdsourcing (Mavridis et al. 2019) and have shown that they can enhance worker engagement during microtask execution (Qiu, Gadiraju, and Bozzon 2020b). In the healthcare sector, studies have shown that chatbots can effectively deliver psycho-education and can treat mental illnesses (Elmasri and Maeder 2016; Fitzpatrick, Darcy, and Vierhile 2017). As of yet, the efficacy of chatbots in training workers for complex therapeutic tasks has not been explored. We aim to investigate the effectiveness of a rule-based chatbot for training workers to deliver emotional support. To this end,

we built *Trainbot*, a conversational interface that leverages Motivational Interviewing (MI) theory, which is a powerful counseling approach for treating anxiety, depression, and other mental problems (Miller and Rollnick 2012).

To study the efficacy of conversational interfaces for crowd training, we conducted a between-subjects experiment on the Prolific crowdsourcing platform. One group of workers (N=100) was trained through a conversational interface (hereafter: Trainbot), whereas the other group (N=100) through a conventional web interface. Both training workflows were identically designed based on MI. The training objective was to prepare workers for coaching a person experiencing stress. It is important to note that we taught MI principles to workers in both Trainbot (treatment group) and simple web interface (control group) and did not treat MI as an experimental parameter. Instead, we wanted to explore the efficacy of a conversational interface for training and deploying non-expert crowd workers for providing emotional support to people in need.

Following the training, we tasked workers with exercising their newly acquired skills with a virtual stressed person in need of support. This virtual person was implemented as a chatbot, which we based on an actual dialogue related to stress management between a user and a robot from prior work (Abbas et al. 2020). We assessed the efficacy of the training by measuring (i) workers' self-efficacy scores before and after the training task; (ii) workers' self-reported scores on enjoyment and stress after the training task; (iii) the number of retakes/attempts to complete quizzes during the training task; (iv) the average number of words used and time spent in answering open-ended questions both during the training and the actual task; (v) two professional clinical psychologists, independently rating the resulting dialogues that workers had with the stressed user, assesssing the workers' effectiveness as coaches for stress management.

We found that workers in the Trainbot group: 1) felt less pressure, retook fewer quizzes, wrote more words and spent more time than the control group; 2) provided psychological interventions that were rated consistently higher by psychologists than the control group; 3) felt a higher self-efficacy in helping deal with stress management after the training task.

Background and Related Work

Motivational Interviewing

Motivational Interviewing (MI) is defined as *"a collaborative conversation style for strengthening a person's own motivation and commitment to change"* (Miller and Rollnick 2012). MI is a powerful counseling approach, which was originally introduced for treating drug addiction and substance abuse (Rollnick and Miller 1995). Nevertheless, recent studies have shown that MI is also effective in treating anxiety, depression, and other mental problems (Westra, Aviram, and Doell 2011). MI uses four fundamental processes to help a therapist support a patient: engaging, focusing, evoking and planning (Miller and Rollnick 2012). In *engaging*, the therapist builds rapport with patients and tries to understand what is going on in their life. In *focusing*, the therapist asks patients to detail their problems, possibly having them identify an inner struggle. In *evoking*, the therapist explores the main reasons for the change. In *planning*, the therapist helps the patient in coming up with their own ideas or action plans for change. In the past, researchers have developed automated systems for health behavior change based on MI techniques (Shingleton and Palfai 2016). In this paper, we explore how a conversational interface can be used to train crowd workers to perform motivational interviewing as stress management coaches.

Affective Crowdsourcing

Within the scope of affective crowdsourcing, collective intelligence has been invoked to deliver complex therapeutic tasks on demand (Morris 2011). For instance, Student Spill[1] and Emotional Bag Check[2] are two emotional support tools that rely on a cohort of trained volunteers to give therapeutic support to students and others. Panoply (Morris, Schueller, and Picard 2015) is a crowd-powered system that leverages crowdsourcing to provide on-demand emotional support. When compared to an online expressive writing group who did not receive support from the crowd, users of Panoply showed higher levels of engagement (Morris, Schueller, and Picard 2015). Similarly, researchers have developed several peer-to-peer online emotional support tools for mental health problems (see review (Ali et al. 2015)).

Panoply employed the cognitive reappraisal technique for training, which involves reframing the meaning of a distorted thought or situation such as irrational or mal-adaptive thoughts (e.g., "I will never pass this exam"). Workers were trained to show empathy, recognize distortions in the input stressors and reframe the distortions. On the contrary, we employed a holistic approach based on MI to train workers for administering an entire therapeutic session -starting from greeting to wrapping up the discussion.

Single Session Therapy

This work also resonates with single session therapy (SST), where each session is treated as the only self-contained session (Bloom 2001). SST is based on the fact that in the majority of cases, a single session of therapy can lead to an overall improvement in the clients (Rosenbaum, Hoyt, and Talmon 1990). As follow up steps, clients are urged to adhere to the positive interventions discussed in the session based on their strengths. STT has been successfully employed to treat mental health problems in children and adolescents (Perkins 2006) and has also helped to significantly lower alcohol use among heavy drinking college students (Samson and Tanner-Smith 2015).

Training and Learning in Crowdsourcing

Researchers have developed several training methods to enhance the performance of unskilled workers for a variety of microtasks. (Gadiraju, Fetahu, and Kawase 2015) exploited the notion of implicit and explicit training in four well-known microtasks on the CrowdFlower platform. In implicit training, workers were only prompted for training

[1] http://www.badgerspill.com/

[2] https://emotionalbaggagecheck.com/

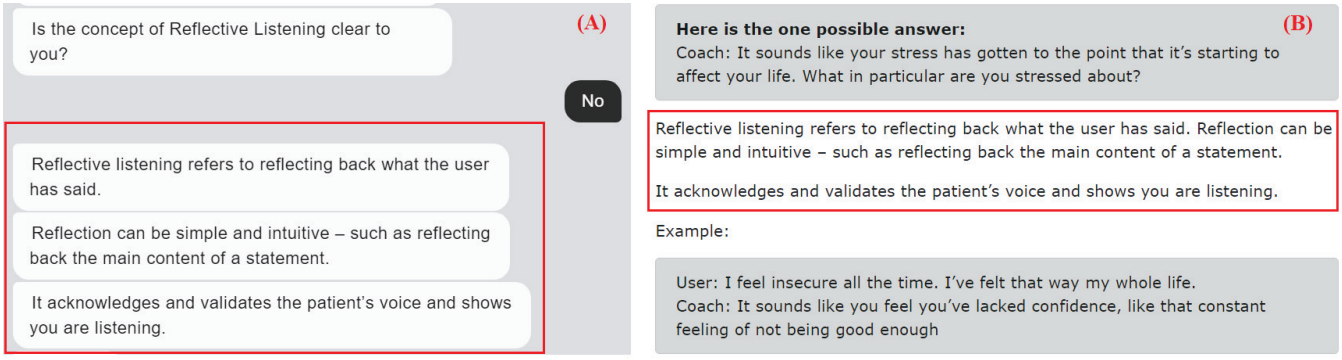

Figure 1: (**A**) Trainbot's interface: after explaining a topic, Trainbot corroborates about the clarity of a topic from workers and provides them with elaborate explanations where required based on the dialogue; (**B**) In the conventional web interface, elaborate explanations are an integral feature of the general topic descriptions.

when they produced flawed output. Whereas in the explicit training, workers completed a training phase before executing the actual tasks. Overall, both forms of training helped workers to improve their performance and aided in the removal of unreliable workers. In another study, researchers compared different training techniques on a complex web-search task. They showed that providing expert examples and asking workers to validate the contributions of peer workers are effective forms of training (Doroudi et al. 2016). (Dow et al. 2012) investigated self-assessment and external assessment for a content creation task on MTurk. In self-assessment, workers reviewed and edited their own work, while in external assessment, workers received expert feedback on their work. Both forms of assessments led to improvements in work quality. Other approaches to training include interactive tutorials (Dontcheva et al. 2014) and priming techniques (Morris, Dontcheva, and Gerber 2012; Gadiraju and Dietze 2017). However, the potential of CUIs for training crowd workers to deliver positive psychological interventions has remained unexplored.

Crowd-Powered Conversational Interfaces

Crowd-powered conversational assistants have been developed to support a variety of tasks. For instance, Chorus (Lasecki et al. 2013) is a text-based chatbot that assists end-users with information retrieval tasks by conversing with a synchronous group of workers. To automate conversation, Evorus (Huang, Chang, and Bigham 2018) builds on Chorus and employs both machine learning and human computation thus enabling a group of crowd workers to collaborate with chatbots. More recently, (Mavridis et al. 2019) investigated the suitability of conversational interfaces for microtask crowdsourcing. They showed that crowd workers perform microtasks more effectively when they interact with a text-based chatbot, compared to the traditional web interface in a variety of typical microtasks (e.g., sentiment analysis, image labeling). It was found that crowd workers exhibited an overall satisfaction while working with the chatbot, and the results produced were of a better quality compared to the web interface (Mavridis et al. 2019). Others studied the impact of different conversational styles employed in a text-based conversational agent on the worker performance and engagement, and proposed models to estimate the conversational styles of workers (Qiu, Gadiraju, and Bozzon 2020a; 2020b). Results indicated that conversational agents with different conversational styles did not impact the output quality, but they had positive effects on worker engagement and worker retention.

Building upon the work of (Mavridis et al. 2019), we studied the extent to which a text-based conversational interface (CUI) can support workers' training, to prepare them for administering therapeutic tasks.

Method

Our aim was to determine whether a text-based conversational interface is more effective and better perceived when compared to a conventional, text-based web interface to train workers on MI for stress management tasks. Thus, we developed two systems, 1) a web interface (control condition), in which we simply detailed MI-based instructions in a conventional way; 2) a conversational interface (Fig.3.a, b), which delivers the same instructions in a conversational style. To safeguard the validity of the comparison, the instructions in both modalities were the same. We developed a rule-based conversational interface, *Trainbot*, based on chat-bubble[3] and Flask. Trainbot also displays a progress bar to inform workers on their progress and the bonuses earned.

Workflow of Trainbot

1. After a worker initiates the training, Trainbot greets the worker and briefly describes the training structure. We structured Trainbot based on the following MI topics: 1) greeting and opening the conversation; 2) reflective listening; 3) showing empathy; 4) asking open questions; 5) affirming the user's strengths/coping skills; 6) wrapping-up the conversation (by encouraging the stressed user to practice some interventions).
2. Next, Trainbot sequentially trains workers on these topics. During the training, Trainbot periodically prompts

[3]https://github.com/dmitrizzle/chat-bubble

Figure 2: Stages involved in the procedure.

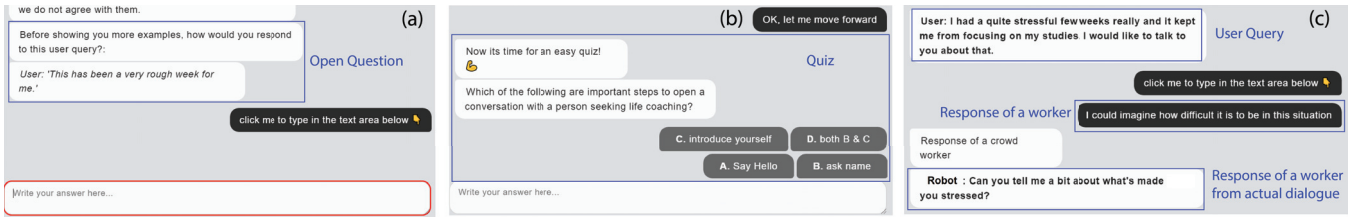

Figure 3: Examples of (a) Trainbot prompting the worker to answer an open-ended question; (b) Trainbot showing a quiz to worker; Test Task interface: on the top, a user query is shown, then a worker can reply by clicking the dark gray bubble. After the worker has replied, a robot's (crowd-powered) response based on a past dialogue is shown.

workers to answer open-ended questions (7 in total; see Fig. 3.a). The purpose of these questions was twofold – to keep workers attentive during training, and to assess workers' engagement with the training by analyzing their responses and the time they would devote to answering those questions.

3. After training workers on a specific topic, Trainbot confirms by asking "did you understand the topic?" and if a worker answers in the negative, Trainbot provides an elaborate explanation with more examples. In the web interface condition, the elaborate explanations are seamlessly included in the topic's descriptions (see Fig. 1). In other words, both interfaces include 'elaborate explanations' to make sure that the workers learn about the topic.
4. At the end of a topic, Trainbot presents workers with short quizzes to solve (5 in total; see Fig. 3.b). Each quiz contains one question. Upon answering a question correctly, Trainbot continues to a new topic. To ensure that workers fully grasp the concepts of MI and understand the instructions, we only allow them to move forward when they complete a quiz correctly. If they fail to answer a question, Trainbot presents them with two options – to either retake the quiz or read the instructions again. The reader can directly experience interaction with the Trainbot. [4]

Conventional Web Interface

The conventional web interface was designed following the progressive disclosure pattern (Nielsen 2006). The instructions could be provided on a single HTML page, but we purposely decompose instructions in several HTML pages to reduce the cognitive load of workers. Each web page pertained to a single topic followed by a new page containing a corresponding quiz. We progressively showed more instructions as the workers proceeded by simply displaying sections of the webpage that were previously hidden.

[4]https://trainbot1.herokuapp.com/training

Participants

We recruited 200 workers (100 for each condition) from the Prolific.ac crowdsourcing platform. We restricted the experiment to only US and UK workers since our task required proficiency in English. We dropped one worker from each condition due to missing data, resulting in a total of 198 workers. Out of 198 unique workers, 60% were female, 39% were male, and 1% did not disclose their gender. 82.5% of the workers were from the UK and the rest were from USA. Their average age was 33.6 years old (SD=11.73). Each worker was paid £3.15 fixed amount (£7.56/h). At the time of writing this paper, this hourly wage was categorized as "good" by Prolific's calculator for both US and UK participants. Workers who participated in one condition were not allowed to participate in the other condition using Prolific's built-in screening feature.

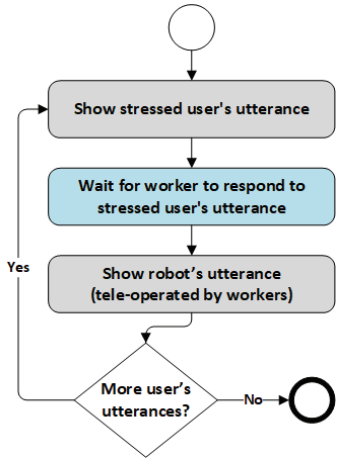

Figure 4: Conversational flow of the test task. The grey-colored rectangles show the actual dialogue between a user and a robot from a prior study in stress management.

Procedure & Measures

Fig. 2 shows the stages involved in the procedure. In this section, we will briefly explain each stage.

(1) Helping Self-Efficacy (Pre-Task) Workers who accepted the task were first asked to fill out the self-efficacy questionnaire regarding their ability to chat with a stressed person. For that reason we used the Session Management subscale (Lent, Hill, and Hoffman 2003). This scale was previously employed to measure the basic helping skills of student helpers who were not trained counselors (Lannin et al. 2019) but had basic communications skills and knowledge to respond to a student's immediate concerns. It consists of 10 items in which we asked crowd workers to rate their confidence to perform specific tasks (e.g. "Help your client to explore his or her thoughts, feelings, and actions."; *0 = no confidence*, 9 = complete confidence). The questionnaire had a high level of internal consistency (Cronbach's $\alpha = .975$). The overall helping self-efficacy score was computed by averaging the scores for the 10 items.

(2) Training Task Next, workers were either redirected to the Trainbot condition or the control condition where they engaged with actual training. The performance in the training was compared between the two groups by calculating the average number of words they composed and the total time they spent (in seconds) in answering open-ended questions. Additionally, we compared the average number of quiz retakes in both conditions.

(3) Enjoyment/Pressure Scales Soon after the training, workers were asked to fill out the Intrinsic Motivation Inventory (IMI) survey (McAuley, Duncan, and Tammen 1989; Ryan 1982), which measures the participants' subjective experience related to a target activity. We administered two subscales from the IMI scale: 1) *Interest/Enjoyment* subscale: It consists of 7 items, which measure the intrinsic motivation of performing the activity on a 7-point scale (e.g. "I enjoyed doing this activity very much"; *1= not at all true*, *7= very true)*; 2) *Pressure/Tension*: It consists of 5 items, which measure how much pressure and tension participants felt while doing an activity (e.g. "I felt very tense while doing this activity"; *1= not at all true*, *7= very true*).

(4) Helping Self-Efficacy (Post-Task) Subsequently, workers were asked to assess their skills using the self-efficacy scale. We sought to find whether their confidence in their helping skills improved after the training.

User:"I moved from Belgium to the Netherlands 3 months ago to do a minor in Industrial design so it's totally a new city ..."
Coach:"Are there any language barriers you have to deal with?"
User:"Not really because in Belgium we speak Flounderish and in Holland, they speak Dutch and it's the same language."

(5) Test Task Workers were then redirected to the test task. The goal of the test task was to evaluate how effective the training was in each modality. We chose a real dialogue from a prior study (Abbas et al. 2020) between a user and a robot related to stress management (the robot utterances were actually written by crowd workers who were teleoperating the robot). The chosen dialogue was simulated by a chatbot. We asked workers in the task to respond to the utterances based on the skills they learned in their training. We implemented the following sequence in the simulated chatbot (Fig. 3.c & Fig. 4): 1) First, the chatbot displays the stressed user's utterance to workers; 2) The chatbot then requests the workers to respond to the user's utterance based on their acquired skills; 3) After that, the response of the robot from the actual dialogue is shown to let the workers know about the context of the conversation (workers were informed that robot's utterances were powered by workers and they were not generated by the robot itself). We showed the response of the robot (acting as a life coach) to prevent confusion about the stressed user's transition from one topic (moving to a new city) to the next (language barrier). The chatbot then repeats these three steps until all the user' utterances from the original dialogue have been shown (Fig 4). Note that we did not receive any criticism from workers in our task about revealing that the robot was a worker in a pre-selected dialogue.

We recruited two clinical psychologists on Upwork.com, experienced in life coaching skills, to evaluate the workers' performance. We paid $80 to each expert. They evaluated the performance of workers on a 7-point scale (*1: totally disagree*, *7: totally agree*) based on the following items:

- ★ The worker's responses show sympathy to the user's situation.
- ★ The worker's open questions help to explore the user's inner struggle.
- ★ The worker's responses reflect and validate the user's statements.
- ★ The worker's proposed solutions are genuine and based on user's strengths and coping skills.
- ★ Please rate the overall performance of workers as a coach for stress management (*1: highly unprofessional* to *7: highly professional*)

For evaluation, we randomly sampled 18 cases from each condition based on an effect size of 0.5 and power of 0.8 (calculated with GPower). However, we discarded one case from each condition due to a duplicate entry, resulting in 17 cases for each condition. As an additional measure, we also calculated the average number of words composed by workers and the average amount of time spent in responding to the stressed user's utterances.

Results

Helping Self-Efficacy (Pre/Post Task)

A two-way analysis of variance (ANOVA) was conducted to study the combined influence of *interventions* (Trainbot

versus control) and *time* (pre-training versus post-training) on helping self-efficacy (HSE). As shown in Figure 5, the main effect of *time* on HSE was significant, $F(1, 392) = 57.68, p < .001$, Trainbot (pre: (M=5.61, SD=1.77), post: (M=7.10, SD=1.36)), control (pre: (M=5.65, SD=1.83), post: (M=6.71, SD=1.67)). We did not find the main effect of *interventions* on HSE, $F(1, 392) = 1.12, p = .29$. We also did not find the interaction effect (*interventions* x *time*) on the HSE scores, $F(1, 392) = 1.64, p = .20$. This shows that both forms of modalities help to significantly improve the workers' confidence about their helping skills.

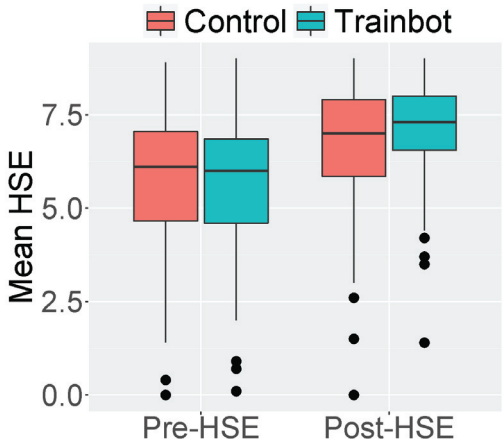

Figure 5: A significant difference between Pre- (helping skills before commencing the training) and Post- (after accomplishing the training) helping self-efficacy (HSE) was observed in both MI-based interventions

Next, we determine the difference in the post-HSE (self-efficacy score after the training task) among the two conditions using independent samples T-test. The post-HSE score corresponding to the Trainbot condition was slightly higher than the control condition. Nevertheless, this difference was not statistically significant, Trainbot (post): (M=7.10, SD=1.36), Control (post): (M=6.71, SD=1.67); $t(196) = 1.806, p = .072$.

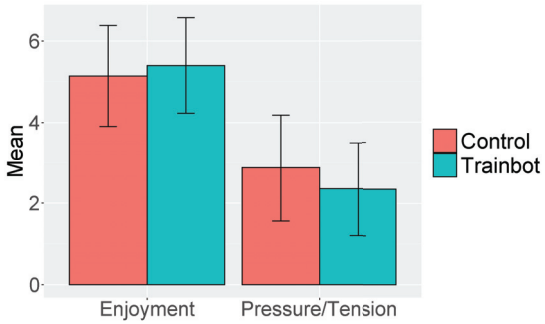

Figure 6: Workers in the CI condition felt less pressure than the control. Workers in the CI condition felt more enjoyment in performing the task than the control condition (although not statistically significant difference).

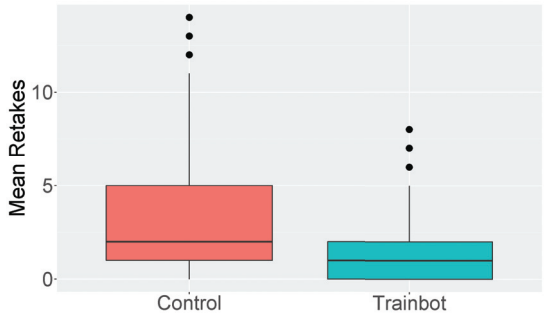

Figure 7: The workers in the Trainbot condition significantly took less retakes than the control group during the training.

Enjoyment/Pressure

The workers' self-reported Pressure/Tension scores showed that workers in the Trainbot condition felt significantly less pressured and tense than the control condition: Trainbot (M=2.35, SD=1.15), control (M=2.87, SD=1.31), $t(196) = -3.018, p = .003$ (Fig. 6).

Regarding the enjoyment scores, workers in the Trainbot condition felt more enjoyment in performing training task than the control group, though this difference was not statistically significant; Trainbot: (M=5.40, SD=1.17), control (M=5.14, SD=1.24), $t(196) = 1.50, p = .135$ (Fig. 6).

Workers' Performance in the Training Task

Importantly, the workers in the Trainbot condition took significantly fewer retakes than the control group; Trainbot (M=1.46, SD=1.70), control (M=3.16, SD=3.181), $t(151) = -4.668, p < .001$ (see Fig. 7).

The average number of words did not differ among the two groups; Trainbot (M=13.39, SD=4.03), Control (M=13.87, SD=6.21), $t(168.1) = -.650, p = .517$. We also did not find a significant difference in the average time they spent; Trainbot (M=103.17, SD=234.38), Control (M=59.70, SD=100.35), $t(132.8) = 1.696, p = .092$.

Workers' Performance in the Testing Task

As we stated in the method section, we asked two professional psychologists to assess how effectively the crowd was able to provide positive psychological interventions to the user in the testing task. As shown in Table 1, we did not find any significant difference in the mean scores for all dimensions (sympathy, open questions, reflective listening, proposing solutions, overall rating). Nevertheless, the psychological interventions provided by workers in the Trainbot condition were rated consistently higher than the control condition across all dimensions. Table 2 presents some examples of positive psychological interventions provided by workers in the testing task from both groups.

A fair level of agreement was found between two raters according to Cicchetti (Cicchetti 1994) (ratings between 0.40 and 0.59 are considered fair). The average measure ICC was .49 with a 95% confidence interval from .335 to .611 ($F(215, 215) = 1.965, p < .001$). We also found a positive moderate correlation between two raters (**r**= .335,

Table 1: T-test results including mean and standard deviation values of the two groups. SYM: sympathy, OQ: open questions, RL: reflective listening, PS: providing solutions

	Trainbot	Control	T-test and Effect size (d)
SYM	$5.63 \pm .910$	$5.38 \pm .857$	$t(32) = .824, p = .41, d = 0.28$
OQ	5.10 ± 1.01	4.88 ± 1.07	$t(32) = .618, p = .54, d = 0.21$
RL	5.06 ± 1.07	4.74 ± 1.13	$t(32) = .854, p = .39, d = 0.29$
PS	4.82 ± 1.01	$4.73 \pm .850$	$t(32) = .275, p = .78, d = 0.09$
Overall	$5.01 \pm .994$	$4.85 \pm .996$	$t(32) = .474, p = .63, d = 0.16$

$p < .001$). We observed little variance across workers in the evaluation by clinical psychologists. Additionally, the performance of workers in the evaluation task was also compared among the two conditions by calculating the average number of words and total time they spent (in seconds) in answering open-ended questions. Though insignificant, the workers in the Trainbot condition wrote more words (Trainbot (M=16.91, SD=6.24), Control (M=16.44, SD=7.28), $t(196) = .484, p = .629$) and spent more time than its counterpart (Trainbot (M=77.18, SD=137.7), Control (M=62.52, SD=99.86), $t(196) = .858, p = .392$).

Discussion

In this paper, we studied the efficacy of a conversational interface (CUI) to train crowd workers for therapeutic tasks. Our results indicate that both forms of interventions resulted in improved post self-efficacy scores. Additionally, workers in the Trainbot condition showed better performance with respect to intrinsic motivation and behavioural measures. In a follow-up evaluation task, professional psychologists rated the performance of workers based on MI, with slightly higher ratings for the treatment group across all dimensions.

Training Systems for Therapeutic Tasks Rely on MI as a Guiding Framework

A significant difference between the pre- and post HSE scores indicates that training crowd workers based on the theoretical framework of MI improved their self-efficacy about their counselling skills regardless of which interface they used. Prior research in educational psychology has shown that self-efficacy is a useful tool for accurately predicting motivation and learning outcomes (Zimmerman, Bandura, and Martinez-Pons 1992). We recommend that future crowd training systems for therapeutic tasks should be designed by using MI as the guiding framework.

However, we did not observe significant differences in the post-training HSE scores between the two groups. This did not confirm our expectations based on a prior study (Fryer, Nakao, and Thompson 2019) where the effectiveness of a chatbot for second language learning was examined. Results from the study indicate that the students who undertook training with the chatbot "learned more" than those with a language learning partner. The learners' qualitative feedback revealed that this result was associated with the fact that the chatbot offered more practice questions and vocabulary than the human partner. Similarly, Trainbot offered elaborate explanations when workers did not understand the topic at hand and presented quizzes and open-questions to test the workers' skills.

A CUI Is Perceived as Being Less Stressful to Interact with for Training

We were wary of the potential of workers getting upset or anxious when trained for mentally demanding therapeutic skills. Workers' self-reported scores indicate that they felt less stress with Trainbot compared to the Control condition. Prior works have attempted to study the efficacy of chatbots for 'delivering' psychological interventions to alleviate stress (Elmasri and Maeder 2016; Fitzpatrick, Darcy, and Vierhile 2017), but as yet, we are not aware of any prior research reporting that "learning" how to conduct psychological interventions through chatbots could also reduce stress and fatigue of the crowd helpers. A possible explanation for this result is that chatbots have been shown to be effective in improving workers' performance and engagement while performing microtasks (Qiu, Gadiraju, and Bozzon 2020b). Another reason could be the engaging elements (emojis, animated GIFs) in Trainbot – it provided encouraging feedback with emojis and animated GIFs when a worker answered a quiz correctly or completed a topic successfully. Emojis and animated GIFs are elements that are commonly used in textual conversational exchanges by users in the real-world (Riordan 2017). A prior study in the mental well-being domain has shown that users evaluated the interaction with a chatbot that was equipped with emojis more positively compared to a chatbot with plain text (Fadhil et al. 2018).

The fact that workers' perceived enjoyment did not differ among the two interventions is congruent with prior research (Kim, Lee, and Gweon 2019); it was found that using a formal conversational agent for conducting interviews against online web surveys resulted in higher-quality data but did not increase the enjoyment of their participants. In our study, the absence of a statistically significant difference for "enjoyment" can be related to the fact that we structured the written instructions and the language style in both modalities in the same way. In a future study, we intend to examine this difference by incorporating a simpler intervention (i.e. a short instructing text, or a Wikipedia article) in comparison to the MI-based intervention that we designed in this study.

Workers in Trainbot Group Made Fewer Mistakes

The fact that the Trainbot group made fewer mistakes while solving quizzes demonstrates the efficacy of the conversational interface for training crowdworkers. Furthermore, workers in the Trainbot group felt more confident for their acquired knowledge about MI. A possible explanation for this result is that chatbot training has been shown to be effective in recalling acquired information and can improve students' learning abilities, though in a pedagogical setting (Abbasi and Kazi 2014). Workers who accomplished training with the Trainbot may have felt more confident when applying the acquired knowledge. This result may also be explained by the reasoning that the conversational interface is more congruent to the actual task. We speculate that this interaction modality congruence might have better prepared workers to complete the requested tasks.

Table 2: Examples of Workers' responses in response to user's utterances in the testing task.

User's Utterance	Worker's response (Trainbot)	Worker's response (Control)	Topic
I had a quite stressful few weeks really and it kept me from focusing on my studies. I would like to talk to you about that.	I'm sorry to hear you'd struggled with your studies; can you outline what's been affecting you this week?	I understand that you've had a few stressful weeks and would like to ask what has been stressful for you in particular?	Greeting, Empathy
I moved from Belgium to the Netherlands 3 months ago to do minor in Textile Engineering so it's totally new city and new the house where I am living in and the subjects are so different...it's is very difficult to find ground to walk on almost.	I can understand why that would make you feel stressed, you've gone through quite a few changes recently. How have you been managing with everything so far?	Having moved to the Netherlands so recently it seems natural you'd have challenges settling in. Have you considered attempting to focus on your strengths whilst adapting to this new teaching style?	Reflective listening, open question
It would be nice to find free time although I'm already behind on my schedule. I should actually spend even more time on my study so that I will be able to finish it on time.	I think you need to not be too hard on yourself! Sometimes having even just half an hour of relax time can help clear your mind and you may find that it makes you more productive afterwards!	It's important to find some free time so that you can better digest what you've learnt. While spending time during this transition on studies is important, it's even more important that you find some free time to relax a little.	Positive interventions
that's a good tip. I did try that...So, I have to go now.. thank you so much for having this conversation with me.	You are more than welcome. I hope your stress eases soon. I am always here if you need somebody to speak to.	I'm happy to have had the opportunity to meet with you and to discuss your current situation.	Closing

Both Interventions Rated Positively by Experts

Experts' ratings for the testing task between Trainbot and the Control group were quite similar for all dimensions of MI. These results mirrored the findings from a previous study (Mavridis et al. 2019; Qiu, Gadiraju, and Bozzon 2020b) where researchers did not find difference in the output quality yielded, regardless of whether a conversational or web interface was used. This shows that MI seems is effective (given that mean scores for both groups are above 4.7), but having a different interaction/modality did not reveal an apparent difference. This is an important outcome for the designers of the affective-crowdsourcing tasks to incorporate the fundamental processes of MI while formulating teaching methods for workers or structuring the instruction manual.

For the Trainbot group, we observed 4.54% improvement for *sympathy*, 6.53% for *reflective listening*, 6.34% for *open questions*, 1.88% for *providing solutions* and 3.24% *overall*. Furthermore, the mean score for the Trainbot group was above 5.0 in all dimensions except for *providing solutions*. Once again this can be explained due to the congruence of the interaction modality between the training phase and the task. Since Trainbot teaches MI through a conversational interface, workers may have felt more confident in applying their acquired skills in the testing task, which was also within a CUI. On the other hand, workers in the Control condition may have felt a higher cognitive dissonance (Festinger 1957) due to the mismatch/conflict between their actual goal (delivering MI via a chat interface) and the medium of learning (via textual instructions).

Differences in the Behavioral Measures for the Training and Testing Task

As we mentioned before, we analyzed worker behavior through the number of words and time spent in both tasks. Although we did not find significant differences in the length of responses, workers in the Trainbot group spent more time in answering open-ended questions in the training task. This may be either due to workers in the Trainbot group taking more time to write messages that would be more impactful, or simply due to finding it harder to do write responses. We plan to carry out a detailed conversational analysis to understand this in the future. We asked workers to answer as quickly as possible after the user's utterance and not to think too long, to reflect the nature of real-time crowdsourcing where workers have limited time to respond (usually few seconds) after the user's utterance (Lasecki, Homan, and Bigham 2014). This can explain our findings regarding the behavioral measures in the testing task for both conditions.

Accountable Crowd-Powered Counselling

Due to the nature of the task, we hoped to instill intrinsic responsibility and accountability among workers to provide valid and respectful counseling. However, given the behavioral dynamics of workers in crowdsourcing marketplaces, one can expect to receive inappropriate or meaningless responses. To increase accountability, one can augment Trainbot with external support from medical professionals, lay persons, and students for vetting crowd-generated responses – a concept known as supportive accountability (Mohr, Cuijpers, and Lehman 2011). However, this solution is not affordable and scalable (Morris, Schueller, and Picard

2015). Another solution is to employ a voting mechanism (Lasecki, Homan, and Bigham 2014) to filter out inappropriate or poorly structured responses by asking additional set of workers to rate the responses. This has been successfully employed in real time crowd-powered conversational agents such as Chorus (Lasecki et al. 2013) for filtering out poor quality answers. To increase accountability, in our future work we will investigate methods to align worker incentives with the incentives and risks posed by the task.

Caveats and Limitations

In our study, we did not consider a non-MI based intervention which can be useful to tease out the impact of training workers on MI as opposed to alternative training methods. Due to privacy and ethical considerations, we used a simulated dialogue from previous work as opposed to employing real users under stress. Future work should evaluate the performance of workers recruited from crowdsourcing marketplaces to deliver on-demand stress management therapy under real conditions and time constraints.

Conclusions & Future Work

Stress is one of the main contributors to mental health problems around the world. The ongoing pandemic is continually affecting the well-being of millions of people. Although AI-based self-help interventions are important to tackle stress, they are limited in their capacity to fulfil wide-ranging emotional needs. Crowd-powered solutions can help to address the immediate concerns of affected people. However, invoking an unskilled pool of workers to deliver positive psychological support can be detrimental. In this paper, we compared training workers with a conversational and a conventional web interface, explored their confidence regarding the skills they acquired, and analysed their performance in test tasks. We designed and developed both systems to train workers on *motivational interviewing* (MI). Although experts' ratings of the workers' life coaching did not reveal significant differences across the two conditions, the mean scores for the Trainbot were consistently higher. Furthermore, workers in the Trainbot group yielded better performance in the training task in terms of making fewer mistakes in answering quizzes and perceiving lesser stress.

In our future work, we plan to study the effectiveness of Trainbot in-the-wild by training workers on-the-fly and then utilizing the trained pool of workers to deliver life coaching in a more realistic context and with a more diverse demographic for participants. On the one hand, this would increase the requirements put upon the training program but on the other, it would capitalize on the versatility of crowdsourcing compared to AI based solutions, in addressing diverse contexts.

References

Abbas, T.; Khan, V.-J.; Gadiraju, U.; Barakova, E.; and Markopoulos, P. 2020. Crowd of oz: a crowd-powered social robotics system for stress management. *Sensors* 20(2):569.

Abbasi, S., and Kazi, H. 2014. Measuring effectiveness of learning chatbot systems on student's learning outcome and memory retention. *Asian Journal of Applied Science and Engineering* 3(2):251–260.

Ali, K.; Farrer, L.; Gulliver, A.; and Griffiths, K. M. 2015. Online peer-to-peer support for young people with mental health problems: a systematic review. *JMIR mental health* 2(2):e19.

Bloom, B. L. 2001. Focused single-session psychotherapy: A review of the clinical and research literature. *Brief Treatment & Crisis Intervention* 1(1).

Cicchetti, D. V. 1994. Guidlines, Criteria, and Rules of Thumb for Evalauting Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment* 6(4):284–290.

Dontcheva, M.; Morris, R. R.; Brandt, J. R.; and Gerber, E. M. 2014. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of CHI'14*, 3379–3388.

Doroudi, S.; Kamar, E.; Brunskill, E.; and Horvitz, E. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of CHI'16*, 2623–2634.

Dow, S.; Kulkarni, A.; Klemmer, S.; and Hartmann, B. 2012. Shepherding the crowd yields better work. In *Proceedings of CSCW'12*, 1013–1022.

Elmasri, D., and Maeder, A. 2016. A conversational agent for an online mental health intervention. In *International Conference on Brain Informatics*, 243–251. Springer.

Fadhil, A.; Schiavo, G.; Wang, Y.; and Yilma, B. A. 2018. The effect of emojis when interacting with conversational interface assisted health coaching system. In *Proceedings of EAI PervasiveHealth'18*, 378–383.

Festinger, L. 1957. *A theory of cognitive dissonance*, volume 2. Stanford university press.

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 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):e19.

Fryer, L. K.; Nakao, K.; and Thompson, A. 2019. Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior* 93:279–289.

Gadiraju, U., and Dietze, S. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 105–114.

Gadiraju, U.; Fetahu, B.; and Kawase, R. 2015. Training workers for improving performance in crowdsourcing microtasks. In *Design for Teaching and Learning in a Networked World*. Springer. 100–114.

Huang, T.-H.; Chang, J. C.; and Bigham, J. P. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proceedings of CHI'18*, 1–13.

Joëls, M.; Pu, Z.; Wiegert, O.; Oitzl, M. S.; and Krugers, H. J. 2006. Learning under stress: how does it work? *Trends in cognitive sciences* 10(4):152–158.

Kim, S.; Lee, J.; and Gweon, G. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of CHI'19*, 1–12.

Lannin, D. G.; Guyll, M.; Cornish, M. A.; Vogel, D. L.; and Madon, S. 2019. The importance of counseling self-efficacy: Physiologic stress in student helpers. *Journal of College Student Psychotherapy* 33(1):14–24.

Lasecki, W. S.; Wesley, R.; Nichols, J.; Kulkarni, A.; Allen, J. F.; and Bigham, J. P. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 151–162.

Lasecki, W. S.; Homan, C.; and Bigham, J. P. 2014. Architecting real-time crowd-powered systems.

Lent, R. W.; Hill, C. E.; and Hoffman, M. A. 2003. Development and validation of the counselor activity self-efficacy scales. *Journal of Counseling Psychology* 50(1):97.

Lucas, G. M.; Gratch, J.; King, A.; and Morency, L.-P. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37:94–100.

Mavridis, P.; Huang, O.; Qiu, S.; Gadiraju, U.; and Bozzon, A. 2019. Chatterbox: Conversational interfaces for microtask crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 243–251.

McAuley, E.; Duncan, T.; and Tammen, V. V. 1989. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60(1):48–58.

Miller, W. R., and Rollnick, S. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Mohr, D.; Cuijpers, P.; and Lehman, K. 2011. Supportive accountability: a model for providing human support to enhance adherence to ehealth interventions. *Journal of medical Internet research* 13(1):e30.

Moore, R. J.; Arar, R.; Ren, G.-J.; and Szymanski, M. H. 2017. Conversational ux design. In *Proceedings of CHI'17*, 492–497.

Morris, R. R., and Picard, R. 2012. Crowdsourcing collective emotional intelligence. *arXiv preprint arXiv:1204.3481*.

Morris, R. R.; Dontcheva, M.; and Gerber, E. M. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16(5):13–19.

Morris, R. R.; Schueller, S. M.; and Picard, R. W. 2015. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of medical Internet research* 17(3):e72.

Morris, R. 2011. Crowdsourcing workshop: the emergence of affective crowdsourcing. In *Proceedings of CHI'11*. ACM.

Morris, R. R. R. 2015. *Crowdsourcing mental health and emotional well-being*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Nielsen, J. 2006. What is Progressive Disclosure?

Perkins, R. 2006. The effectiveness of one session of therapy using a single-session therapy approach for children and adolescents with mental health problems. *Psychology and Psychotherapy: Theory, Research and Practice* 79(2):215–227.

Qiu, S.; Gadiraju, U.; and Bozzon, A. 2020a. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW1):1–23.

Qiu, S.; Gadiraju, U.; and Bozzon, A. 2020b. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of CHI'20*, 1–12.

Riordan, M. A. 2017. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology* 36(5):549–567.

Rollnick, S., and Miller, W. R. 1995. What is motivational interviewing? *Behavioural and cognitive Psychotherapy* 23(4):325–334.

Rosenbaum, R.; Hoyt, M.; and Talmon, M. 1990. The challenge of single-session therapies: creating pivotal moments', in r. wells and v. gianetti (eds), the handbook of brief therapies. new york: Plenum.

Ryan, R. M. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43(3):450.

Samson, J. E., and Tanner-Smith, E. E. 2015. Single-session alcohol interventions for heavy drinking college students: A systematic review and meta-analysis. *Journal of Studies on Alcohol and Drugs* 76(4):530–543.

Shingleton, R. M., and Palfai, T. P. 2016. Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. *Patient education and counseling* 99(1):17–35.

Taelman, J.; Vandeput, S.; Spaepen, A.; and Van Huffel, S. 2009. Influence of mental stress on heart rate and heart rate variability. In *4th European conference of the international federation for medical and biological engineering*, 1366–1369. Springer.

Taylor, S.; Landry, C.; Paluszek, M.; Fergus, T. A.; McKay, D.; and Asmundson, G. J. 2020. Development and initial validation of the covid stress scales. *Journal of Anxiety Disorders* 102232.

Vogel, A. P., and Morgan, A. T. 2009. Factors affecting the quality of sound recording for speech and voice analysis. *International journal of speech-language pathology* 11(6):431–437.

Westra, H. A.; Aviram, A.; and Doell, F. K. 2011. Extending motivational interviewing to the treatment of major mental health problems: current directions and evidence. *The canadian journal of psychiatry* 56(11):643–650.

Zimmerman, B. J.; Bandura, A.; and Martinez-Pons, M. 1992. Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American educational research journal* 29(3):663–676.