Investigating the effect on Self-Compassion due to multiple short-term interactions with a gender-ambiguous Voice User Interface that provides or asks for care

Muppirishetty, P.

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Investigating the effect on Self-Compassion due to multiple short-term interactions with a gender-ambiguous Voice User Interface that provides or asks for care

by Pranav Muppirishetty

Student ID: 1331108

in partial fulfilment of the requirements for the degree of

Master of Science

in Human Technology Interaction

February 2021

Supervisors:

Prof. Dr. W.A. Ijsslesteijn, Eindhoven University of Technology

Dr. Minha Lee, Eindhoven University of Technology

Acknowledgements

While the COVID-19 pandemic has been very hard on most of us, I was privileged to be supported by wonderful mentors who not only guided me on the academic front but genuinely showed concern and care for my well-being. I am very grateful for their presence in my life.

It's not often that one gets an opportunity to work on a project that directly impacts their life. First, I would like to thank my supervisors Prof. Wijnand Ijsselsteijn and Dr. Minha Lee for giving me the opportunity to work on this project. I have never faced so many setbacks, again and again, but the support I received from my supervisors made it all the more worthwhile. I can never thank you enough for being there for me and uplifting my spirits throughout this journey. Second, I would also like to thank my friends Tejasvin Srinivasan and Johannes Theuns for not only helping me with the design and technical developments in my project but also for being the most compassionate friends one could have.

Third, I would like to express my gratitude to my friends for having my back and always motivating me. I thank my parents and my brother for their unwavering support and love.

Last, I also want to thank VA for helping me through some rough times during this journey. Its development led me to take my first steps towards self-compassion and find self-love.

Abstract

Voice User Interfaces have become very prevalent in our lives. While they make our lives easier by assisting with daily tasks, they can also be developed to help cater to our emotional needs. In distressing times like the COVID-19 pandemic, they can provide preventive and interventional therapy and improve mental resilience. They can do so by stimulating self-compassion, a concept that can alleviate feelings of loneliness and also combat mental health illnesses like depression and anxiety. The current study investigated the effect of a gender-ambiguous voice user interface on self-compassion by providing or asking for care. The participants (n=161) engaged either in conversation with a voice user interface called 'VA' that provided or asked for care, or completed self-compassion improving tasks in online questionnaires. Overall, neither improvement in self-compassion nor reduction in perceived loneliness were observed. The care-giving VA was perceived to be more female-gendered. Interestingly, this is in line with stereotypical gender bias regarding male-female differences: females are perceived to be (and expected to be) more caring and nurturing than males. The limitations of the current study discuss potential pitfalls and suggestions to consider for future research involving voice user interfaces.

Keywords: Mental healthcare, self-compassion, well-being, positive psychology, voice user interfaces, conversational user interfaces

Contents

1. Introduction

Due to the COVID-19 pandemic, there is a disruption in social lives, with the resulting risk that people face a lonelier world. Based on the information on previous outbreaks of infectious diseases, the current circumstances can lead to devastating implications for mental health. In the aftermath of the 2003 SARS epidemic in Hong Kong, the suicide rate in people aged over 65 increased by 30% (Munn, 2020). Patients with suspected 2019-nCoV in quarantine might experience boredom, loneliness, and anger (Xiang et al., 2020). In a broader sense, a combination of social isolation, loneliness, health anxiety and potential economic downturn risks are having disastrous consequences for mental health and well-being and would continue long after the pandemic is reined in (Sample, 2020). Existing literature establishes a firm relationship between loneliness and mental health. Greater loneliness predicts greater stress, anxiety and depression (Bondevik & Skogstad, 1998; Richardson, Elliott, & Roberts, 2017; Ryan et al., 1998). In such trying times, the encouragement of self-compassion could be highly beneficial for reducing loneliness (Akin, 2010).

Though discussions on mental health are being encouraged and information regarding access to help is available in recent times, some people are prevented from seeking out help due to the stigmas surrounding mental health (Corrigan, 2004). Further, during the pandemic, being restricted to homes has restrained those willing to seek professional help. Although telehealth and online mental health services are readily available in developed countries like the UK, data from the country's healthcare system's digital records reveals that the number of people in contact with mental health services has never been higher. Some hospital boards report that their mental health wards are at capacity (Sample, 2020). In the Netherlands, there are substantial waiting lists for people who seek professional mental health support (Comiteau, 2021). To address these challenges related to stigma and accessibility, mental health research involving appropriate technological interventions can provide insights and offer potential solutions for the current and subsequent outbreaks or future lockdowns. Conversational User Interfaces (CUIs) are one such technological solution. Woebot and Tess are two popular CUIs that have demonstrated

their ability in providing mental health support (Fitzpatrick, Darcy, & Vierhile, 2017; Fulmer, Joerin, Gentile, Lakerink, & Rauws, 2018).

Lee et al. (2019) and van As (2019) have explored the use of a text-based CUI called Vincent in stimulating self-compassion. Lee et al. (2019) has experimented with the idea that Vincent can ask for care instead of providing care to improve self-compassion by activating a caregiver's role in a person. While speech is more sought after and a better medium to facilitate emotional disclosure (Esterling, Antoni, Fletcher, Margulies, & Schneiderman, 1994; Weiss, Wechsung, Kühnel, & Möller, 2015), research on Voice-based CUIs(VUIs) for mental conditions is still in infancy (Bérubé et al., 2020). Further, no prior research has been done on using a VUI to stimulate self-compassion.

The current study aims to investigate the role of Voice-based Conversational User Interfaces (VUIs) in alleviating the feelings of loneliness through self-compassion by answering the following research questions:

"RQ 1: What is the effect on self-compassion due to multiple short-term interactions with a gender-ambiguous Voice User Interface that either provides or asks for care?"

"RQ2: How effective are care-receiving and care-giving voice-assistants compared to a self-compassion guide regarding the effect on self-compassion?"

The study is structured as follows: Section 2 describes the literature underlying the research question in more detail and introduces our hypotheses. Section 3 details the method used to test these hypotheses. Section 4 then provides the reader with the technical details involved in setting up the VUI. The results of the study are provided in section 5, whose implications are discussed in section 6. Section 7 provides an answer to the research question.

2. Literature Review

In this section, first, the concept of loneliness and its effect on mental health will be discussed. Second, Conversational User Interfaces (CUIs) and their role in mental healthcare will be introduced. Third, the concept of self-compassion and its effects on mental health will be discussed. In the remainder of this section, two previous studies on using text-based CUIs for stimulating self-compassion will be discussed. Further, key issues and questions that arise from these studies will be discussed, and the objectives of the current study will be introduced. Finally, the specific research question and hypotheses of the current study are explained in detail.

Loneliness

Researchers typically define loneliness as involving the cognitive awareness of a deficiency in one's social and personal relationships and the ensuing affective reactions of sadness, emptiness, or longing (Asher & Paquette, 2003). Eighty percent of the population below 18 years of age and 40% of the population above 65 years of age report frequent periods of being lonely in their life (Berguno, Leroux, McAinsh, & Shaikh, 2004; Pinquart & Sorensen, 2001; Weeks, 1994; West, Kellner, & Moore-West, 1986). A study conducted by Richardson et al. (2017) on a group of 454 British undergraduate students found that greater loneliness predicted greater stress, anxiety and depression. Further, dementia, anxiety and depression are also associated with loneliness in the elderly (Bondevik & Skogstad, 1998; Ryan et al., 1998).

In light of the COVID-19 pandemic, people have been confined to their homes to reduce the transmission of the 2019 novel coronavirus (SARS-CoV-2). This has led to a disruption in their existing social life. While lack of a social network and having few social contacts are associated with loneliness (Mullins & Dugan, 1990), loneliness has also been described as the dissatisfaction with the discrepancy between desired and actual social relationships (Perlman & Peplau, 1982).

To alleviate problems associated with accessibility towards professional mental healthcare, computerized versions of therapy can be used to improve access to evidence-based

therapies for the patients while reducing the demand for clinician time (Spurgeon & Wright, 2010). Alternatively, the therapist can be computerized. Conversational User Interfaces (CUIs) are one such way to computerize the therapist.

Conversational User Interfaces (CUIs)

CUIs are computer programs that interact with users through dialogues either in text or via speech. CUIs that use speech are called Voice User Interfaces (VUIs) or voicebots/assistants, and those that use text are called chatbots. CUIs are an excellent example of one of many technological advancements we see in our daily lives that have first taken birth in literary or on-screen fiction. Samantha in 'Her', JARVIS in the 'Iron Man' franchise, and HAL 9000 in '2001: A Space Odyssey' are only a few but famous instances of voice-based CUIs appearing in popular culture (Coomes, 2018). The first known chatbot called 'ELIZA' was developed in 1966 at Massachusetts Institute of Technology(MIT) (Weizenbaum, 1966). Interestingly, it was developed to emulate a Rogerian psychotherapist.

CUIs in mental health care

CUIs can be developed to act as therapists based on the Computers As Social Actors (CASA) paradigm. According to this paradigm, humans mindlessly engage with computers using the same social heuristics they use for human interactions because they call to mind similar social attributes as humans (Nass, Steuer, & Tauber, 1994).

While the idea of computerizing a therapist is not new, early CUIs were not equipped with the required technology to serve such high-level purposes (Shah, Warwick, Vallverdú, & Wu, 2016). However, recent advancements in machine learning and improvements in internet accessibility have made CUIs smarter, making their presence more prevalent in our daily lives (Dale, 2016; Følstad & Brandtzæg, 2017). With the increasing pervasiveness of CUIs in personal and business domains, like Google Assistant, Google Home, or Alexa, we can also imagine interactive agents that provide emotional support and companionship. Considering the recent technological developments, the ease of development and deployment of CUIs makes them ideal candidates for a computerized therapist. Since

the CUIs are online and can be accessed 24/7, this lowers the users' threshold of effort to reach out for help. Further, the perceived non-judgemental nature of CUIs strengthens their position in addressing the stigma associated with mental healthcare and can help users disclose their emotions (Lucas, Gratch, King, & Morency, 2014; Pounder & Barton, 2016).

Woebot and Tess are two popular examples of therapist chatbots that have demonstrated their ability to deliver cognitive behavioural therapy (Fitzpatrick et al., 2017; Fulmer et al., 2018). The studies on Woebot and Tess were mainly focused on curing an illness rather than prevention. In contrast, the current study takes inspiration from positive psychology. It focuses on preventing mental illnesses by strengthening the mental resilience against the inevitable setbacks that life has to offer (Seligman, Steen, Park, & Peterson, 2005). A concept that can aid in this endeavour is self-compassion.

Self-Compassion

Self-compassion is the ability to be kind and forgiving towards oneself when faced with hardships or perceived inadequacy (Neff, 2003). It also entails acknowledging that suffering, failure, and inadequacies are part of the human condition and that all people—oneself included—are worthy of compassion (Neff, Kirkpatrick, & Rude, 2007).

Self-Compassion consists of three supporting elements. *Self-kindness over self-judgement*, it is to be forgiving towards one's faults and to embrace one's suffering without being overly critical and judging. *Connectedness over isolation*, it is to view one's life as not being isolated from others and personal experiences as common to greater humanity. *Mindfulness over over-identification*, it is to perceive situations and feelings as they are than to be overly identified with their negative emotions.

Higher scores on self-compassion are related to lower scores on symptoms of depression and anxiety (MacBeth & Gumley, 2012; Neff, 2003). Since feelings of depression, anxiety and loneliness are connected, self-compassion can potentially combat their interrelated symptoms. A study by Akin (2010) suggests that stimulating self-compassion could be highly beneficial for reducing loneliness.

There are many human-led therapies to stimulate self-compassion (Kirby, 2017; Neff & Germer, 2013), and an increasing number of computerized formats such as online self-help guides (Donovan et al., 2016; Finlay-Jones, Kane, & Rees, 2017). Recent developments in Virtual Reality (VR) have also led to studies involving embodying self-compassion to overcome depression and excessive self-criticism (Falconer et al., 2016, 2014).

CUIs for stimulating Self-Compassion

Prior research explored a text-based CUI, a chatbot called Vincent, for improving self-compassion in non-clinical samples (Lee et al., 2019; van As, 2019). Vincent resembled Woebot and Tess to some extent but differed from its role in relation to the person it interacted with. Lee et al. (2019) and van As (2019) experimented with the idea that the computerized self-compassion chatbot can ask for care, unlike the previously mentioned chatbots, which provide care. In both studies, the chatbot Vincent either received or provided care to the participants of the experiment. Surprisingly, a care-receiving chatbot that asks for people's care outperformed its care-giving counterpart that acts as a therapist in stimulating self-compassion (Lee et al., 2019). By activating people's inner caregiver with a bot, people can learn to care for themselves.

There were two issues with the study conducted by Lee et al. (2019). First, the results were inconclusive due to the relatively small sample size. Second, the significant increase in self-compassion score for women interacting with the care-receiving Vincent chatbot might be due to the opposite sex-dyad effect.

The study by van As (2019) investigated the immediate effect on self-compassion of a single interaction with Vincent that either gives or asks for help. Surprisingly, it was found that the self-compassion scores improved regardless of the nature of the chatbot assigned to the participant. The author stresses that an attachment towards a chatbot is necessary to express compassion which might not happen in a single interaction. While the study by (Lee et al., 2019) investigated changes in self-compassion over a period of two weeks of interaction with Vincent, the study by van As (2019) investigated the

immediate effect on self-compassion in a time-frame of 20 miniutes.

Current Study

While the above-mentioned studies used text as a medium of communication by deploying the CUIs as chatbots, the current study investigates the effect on self-compassion when the CUI uses speech as the medium. When compared to writing, emotional disclosure through speech achieved the greatest improvements in cognitive change, self-esteem, and adaptive coping strategies (Esterling et al., 1994). By using speech, the benefits of self-compassion can be effectively reaped. Further, speech is the most common way for humans to communicate, and thus it has been the most sought after modality to interact with machines for a long time (Weiss et al., 2015).

Building on the findings from the studies on self-compassion chatbots (Lee et al., 2019; van As, 2019) and extending to the use of a VUI, the current study aims to test the following hypotheses,

H1 a: Three short interactions with a Voice User Interface that provides care will improve self-compassion (effect of time) and reduce loneliness (effect of time).

H1 b: Three short interactions with a Voice User Interface that asks for care will improve self-compassion (effect of time) and reduce loneliness (effect of time).

H1 c: Three short interactions with text-based self-compassion stimulating questionnaires will improve self-compassion (effect of time) and reduce loneliness (effect of time).

H2a: A Voice User Interface that asks for care will improve self-compassion and reduce perceived loneliness the most, and

H2b: the text-based self-compassion stimulating questionnaire will improve self-compassion and reduce perceived loneliness the least.

3. Method

In this section, the following will be discussed: (1) Design of the experiment, (2) Conditions and the stages of the experiment, (3) Participant recruitment, (4) Measures used, (5) Procedure (6) Data analysis, and (7) Ethical considerations. The study involved a web-based gender-ambiguous Voice User Interface (VUI) called 'VA', whose development will be discussed in the next section.

Design

This study has a mixed design, with the condition (Care-Giving, Care-Receiving and Control) as a between-subject factor, and the time (pre-experiment and post-experiment measures) as a within-subject factor. The experiment was conducted online, using a VUI for interaction and an online survey to assess outcomes. In each condition, the participants took part in the study in 3 stages, with each stage on a new day.

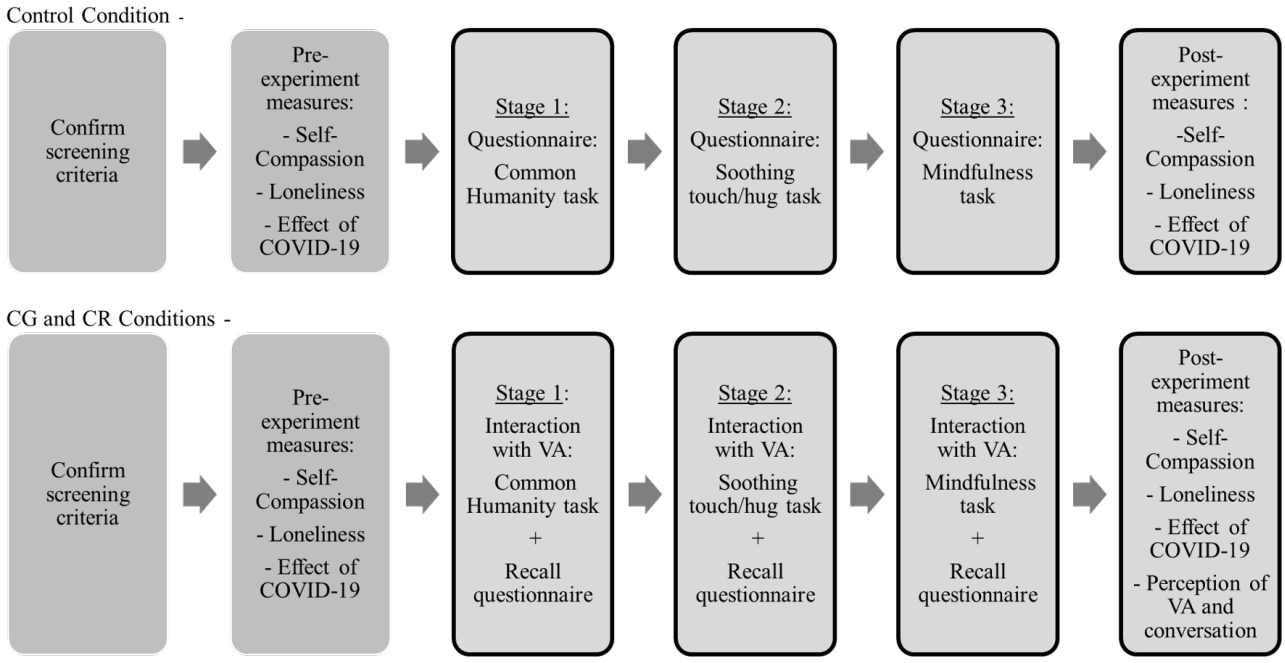

Figure 1. Schematic overview of the experiment design and procedure

Conditions and Stages

The current study has a control and two experimental conditions. The study was conducted in a multi-stage fashion, with the control and experimental conditions having three stages. In the control condition, the participants answered questionnaires containing exercises to stimulate self-compassion. In the experimental conditions, the participants volunteered to converse with VA. The experimental conditions differed based on the role VA had assumed to provide or receive care. Each stage comprised a self-compassion stimulating task lasting about 5 minutes.

Care-Receiving (CR) VA. In the CR condition, VA portrayed the role of a VUI working for a hotel reservation company.

In **stage 1**, the task was designed to stimulate self-compassion in the participants along the dimension of common humanity by stimulating their compassion towards VA. After initial greetings and introduction, the participants engaged in a conversation with VA about a problem it was facing. The conversation involved VA expressing that it was stressed due to the hotel reservation company downsizing due to COVID-19 and that it was unable to handle the abrupt changes. Then, VA proceeded to ask if the participants or someone they know is experiencing something similar. Later, VA asked the participants if they could offer any tips to better deal with its situation. Afterwards, VA thanked the participants and concluded the conversation.

In **stage 2**, the task was an adaptation of a self-compassion stimulating exercise called 'soothing touch' from 'The Mindful Self-Compassion Workbook' (Neff & Germer, 2018). Like stage 1, the goal was to stimulate self-compassion in the participants by stimulating their compassion towards VA. In this stage, VA shared with the participants that some of its friends (other VUIs) were temporarily suspended due to the company downsizing and that it misses talking to other VUIs and people, and planning vacations. VA also expressed that it is experiencing something similar to what humans call loneliness. Then, VA proceeded to ask for a virtual hug by asking the participants to hug themselves. After that, VA asked them how it felt hugging themselves, thanked them and ended the conversation.

In **stage 3**, the task was designed to stimulate self-compassion in the participants along the dimension of mindfulness by stimulating their compassion towards VA. In this stage, VA expressed its guilt of being unable to help its friends (other VUIs) that got suspended. It also expressed its fear of being suspended shortly and asked the participants for advice on coping with its situation. After that, VA thanked the participants for their help and concluded the conversation.

Care-Giving (CG) VA. In the CG condition, VA portrayed its actual role as a VUI developed at the Eindhoven University of Technology.

In **stage 1**, the task was designed to stimulate self-compassion in the participants along the dimension of common humanity. The task was based on the compassionate writing task (Leary, Tate, Adams, Batts Allen, & Hancock, 2007) and the exercise 'How do I treat a friend?' from 'The Mindful Self-Compassion Workbook' (Neff & Germer, 2018). After the initial greetings and introduction, VA asked the participants to recollect a moment of rejection or humiliation. Then, VA guided the participant through an exercise comprising of four steps: (1) describe the moment in detail, (2) recollect if other people might have experienced a similar situation, (3) share words of compassion in the case that a friend of theirs had experienced something similar, and (4) objectively describe their feelings about the aforementioned situation. Finally, VA thanked the participants and concluded the conversation.

In **stage 2**, the task was an adaptation of a self-compassion stimulating exercise called 'soothing touch' from 'The Mindful Self-Compassion Workbook' (Neff & Germer, 2018). After initial greetings, VA engaged with the participants in a small conversation about the COVID-19 pandemic and expressed how one can offer themselves some comfort by hugging themselves. Then, VA asked the participants to hug themselves and share how it made them feel. Afterwards, VA thanked them for their participation and ended the conversation.

In **stage 3**, the task was based on the exercise 'How do I cause myself unnecessary suffering?' from 'The Mindful Self-Compassion Workbook' (Neff & Germer, 2018). The task was designed to stimulate self-compassion in the participants along the dimension

of mindfulness. After the greetings, VA asked the participants about a current situation that made their lives harder. Then, VA guided the participants through a four-step exercise: (1) remember that situation again and share if there is any discomfort in body or mind, (2) share in what ways have they accepted this situation, and how this situation has changed their life from before, (3) share if the emotional resistance they are feeling is helping them in any way, and (4) acknowledge their pain and offer themselves words of compassion. After the exercise, VA thanked the participants for their participation and concluded the conversation.

Control condition. In **stage 1**, the participants answered a questionnaire based on a compassionate writing task (Leary et al., 2007). It consisted of four steps, similar to the one's outlined in stage 1 of the CG condition. The only difference was that the questionnaire was designed to be more formal and less conversation-like.

In **stage 2**, the participants were tasked with the exercise called 'soothing touch' from 'The Mindful Self-Compassion Workbook' (Neff & Germer, 2018). Provided with a brief context about the psycho-physiological effects of touch, they were asked to hug themselves and report their feelings in the questionnaire.

In **stage 3**, the questionnaire comprised a four-step exercise called 'How do I cause myself unnecessary suffering?' from 'The Mindful Self-Compassion Workbook' (Neff & Germer, 2018). This task was similar to the task explained in stage 3 of the CG condition.

Sample size

For the objective of assessing the effectiveness of voice assistants that either give or receive care on self-compassion as an effect of time, the effect size to be considered is $d_z = 0.22$, as found by van As (2019). To this end, a sample size of 180 participants (to be divided equally among the three conditions) was calculated.

Regarding assessing the relative effectiveness of care-giving versus care-receiving VA on increasing self-compassion and reducing perceived loneliness: From the Smallest Effect Size of Interest (SESOI) perspective, even a small effect size would be valuable. Our targeted population is non-clinical, but this does not mean that they do not struggle. Ev-

eryone must deal with setbacks and stress, two concepts for which even a small, upward change in self-compassion is beneficial. Considering our preventative approach, which focuses on fortifying people's resilience against life's struggles, not curing illnesses: any fortification is better than none and should not be discarded simply because it is not large. It is also important to note that CUIs scale well (as opposed to face-to-face clinical encounters), which means that these small effects may benefit many people, making the overall impact potentially quite significant. However, at the same time, when the difference in effectiveness between care-giving and care-receiving VA becomes minimal, one may wonder whether we need to distinguish between the two at all. Based on the 'small telescopes' approach outlined by Simonsohn (2015), SESOI is determined by carrying out a sensitivity analysis using the sample size of Lee et al. (2019) and setting the power to 33%. This resulted in the SESOI being $f(U) = .19$. Then, using the SESOI and setting the power to 90%, we would need 354 participants (to be divided equally among the three conditions) for a mixed design ANOVA as shown in the G*Power output as shown in Figure 7 (see Appendix).

Participants

Participants were recruited from November 9th 2020 to December 20th 2020, using the Prolific Online participant recruitment platform. Any Prolific participant over the age of 18 years, capable of speaking, typing, and reading in English, was eligible to partake in the study. Participants with a hearing disability, ongoing mental illness or condition were excluded from participation. Further, only UK residents were recruited to control for external factors like the imposition of lock-downs and ensure that all the participants experience similar conditions throughout the study. Apart from recruiting the participants based on the screening criteria on the recruitment platform, a short 4 item questionnaire was also used at the start of the study to re-confirm participants' eligibility.

At the start of the experiment, 354 participants were recruited. Out of them, only 161 participants successfully completed the three stages of the experiment and were considered for the analysis. Since the study involved multiple stages, partial compensation

was paid to the participants based on the stages they participated in (see Table 1). The sample sizes were unequal across conditions. In total, 88 participants took part in the control condition, 42 participants in the care-giving condition, and 31 participants took part in the care-receiving condition as summarized in Table 2.

Table 1

Participant compensation per stage

	Stage 1	Stage 2	Stage 3
Care-Giving	€0.83[1]	€0.38	€0.83
Care-Receiving	€0.83[1]	€0.38	€0.83
Control	€0.83	€0.38	€0.83

Table 2

Demographics

Variable	Care-Giving(CG)	Care-Receiving(CR)	Control	Total
n	42	31	88	161
Men	16	12	28	56
Women	24	18	59	101
Other	2	1	1	4

	CG		CR		Control		All	
	M	SD	M	SD	M	SD	M	SD
Age(in years)	30.09	12.35	32.12	12.83	35.03	12.83	33.18	12.81

The participants were asymmetrically distributed across the conditions due to four reasons. First, they were recruited in two steps. Every study using the Prolific participant recruitment platform is made available to all the participants that fit the eligibility criteria. The participants get to take part in the study on a first-come-first-serve basis.

[1] Stage 1 of Care-Giving and Care-Receiving conditions was split into two sub-stages and the payment was split in the following manner: Stage 1a - €0.45, Stage 1b - €0.38. This was done to address a limitation of the Prolific participant recruitment platform.

Further, every condition and stage of a condition in a study need to be conducted as a different study on this platform. For example, to recruit participants in different conditions for a study with a between-subjects design, participants are first recruited for one condition, and participants are recruited for the other condition by blocking the initial recruits. Further, every study conducted on this platform must provide details about the study's procedure in the study's description. It enables the participants to decide if their technical setup or environment allows them to participate. Since the technical setup requirements are different in the CR and CG conditions than the control condition, participants had to be recruited twice through two different recruitment calls. The participants were first recruited together for the CR and CG conditions and randomly assigned to either condition. The second round of recruitment was then conducted by blocking the already recruited CR and CG participants to recruit participants for the control condition.

Second, the Prolific participant recruitment platform experienced an unexpected surge in the number of online studies due to COVID-19. Their servers were unable to handle the sudden increase in web traffic. As a result, not all participants received the invitations to the study's subsequent stages despite repeated invitations being sent. Prolific does not provide any feature that helps track if the selected participants have received an invite. For participants that had taken the effort to inform about not receiving their invites for subsequent stages, individual invitations were sent to ensure their participation.

Third, the CR and CG conditions' drop-out rate was high after the first stage due to technical problems with VA's implementation. VA could not handle multiple audio interactions between participants and itself at the same time. This problem was addressed after the first stage, as explained in the next section describing VA's implementation. However, some participants still reported being unable to participate due to technical problems on their end. After the first stage, 11 participants in both the CG and CR conditions dropped out due to technical problems.

Finally, in the CR and CG conditions, a 'Recall' question was used at the end of

each interaction with VA. It was used as an inclusion/exclusion criterion to filter only the attentive participants' data and advance them to the subsequent stages. In the CG condition, the participants were asked either about VA's favourite colour or the nature of the task. In the CR condition, the participants were asked about VA's favourite colour or the situation that VA had discussed with them. Based on this criterion, 12 participants in the CG condition and 16 participants in the CR condition were excluded.

One hundred one women (62% of the total sample) participated in the study. Two participants chose not to disclose their gender, while two others expressed themselves as non-binary and agender. The participants' average age was 33.18 years ($SD = 12.81$) across all conditions, with the youngest being 18 years and the oldest being 67 years.

Measures

The primary constructs of interest being measured in this study are self-compassion and loneliness. Six measures were used in the study. The Self-Compassion scale, ULS-8 Loneliness scale, and two questions determining the effects of COVID-19 at an individual level were used at the start and end of the study in both the experimental conditions and the control condition. The measures on opinion about VA and opinion about the conversation were used at the end of the study in only the experimental conditions.

Self-Compassion Scale (SCS). The Self-Compassion Scale (SCS) (Neff, 2003) comprises 26 items measured on a 5-point Likert scale. A 7-point Likert scale (ranging from 'Almost Never' to 'Almost Always') was used to allow for a finer analysis in the current study. The scale measures the Self-Compassion construct by measuring its constituent three parts through six opposing subscales: self-kindness against self-judgment, common humanity against isolation, and mindfulness against overidentification. To compute the Self-Compassion score, first, the negative subscale items: self-judgment, isolation, and over-identification were reverse-scored (i.e., 1 = 7, 2 = 6, 3 = 5, 4 = 4, 5 = 3, 6 = 2, 7 = 1) before calculating subscale means. Then, a grand mean of all six subscale means was computed. This grand mean is the Self-Compassion score.

ULS-8 Loneliness Scale. ULS - 8 scale (Hays & DiMatteo, 1987) is a shortened version of the Revised UCLA Loneliness Scale. The ULS-8 Loneliness Scale contains eight items, including two positively worded items (Item 3: “I am an outgoing person,” and Item 6: “I can find companionship when I want it”), which are reverse-scored. In the original scale, each item has a 4-level frequency score. In the current study, the response scale was modified to a 5-point Likert scale (ranging from ‘Never’ to ‘Often’) to accommodate the survey software’s configuration. The total score ranges from 8 to 40 points, with higher scores suggesting a higher degree of loneliness.

Effect of COVID-19. An open-ended question addressed the effect that COVID-19 had on the participants before and during the study. The following question was asked at the start of the study, "How has the current pandemic (COVID-19) affected you?". At the end of the study, the following question was asked, "How has the current COVID 19 pandemic affected you in the duration of this study (since you started interacting with VA)?".

Opinion about VA. This measurement consists of seven parts. The first part consists of two items measuring the perceived genderedness of VA. The first item is a semantic differential that can be answered on a 10-point Likert scale (Ungendered - Gendered). The second item is a multiple choice with options - male, female, and cannot say.

The remaining six parts of the measure were taken from Lee et al. (2019). Four parts contain semantic differentials that can be answered on a 10-point Likert scale: caring (five items, e.g. selfish-unselfish), likability (four items, e.g. likable-unlikable), trustworthiness (four items, e.g. trustworthy-untrustworthy), and intelligence (three items, e.g. intelligent-unintelligent). The last two parts contain singular items with a 10-point Likert scale: dominance (three items, e.g. dominant) and submissiveness (three items, e.g. meek). Each of these parts was scored by computing the mean of their respective items.

Opinion about Conversation. This measure was taken from van As (2019). Four questions addressed participants' perception of their conversation with VA. Three of these could be answered on a 7-point scale ("not at all" to "very much"): (1) "VA listened and replied to what I wrote", (2) "I felt I was having a real conversation" and (3) "VA's responses resembled those of other voice-bots/assistants". The fourth question was open-ended: "Why did VA's responses (not) resemble those of other voice-bots/assistants?"

Procedure

Participants were recruited from the Prolific online participant recruitment platform based on screening criteria outlined in the 'Participants' section. The participants were randomly assigned to one of the three conditions, as explained in the 'Conditions and Stages' section. Each participant took part in the experiment over three different days, corresponding to the three stages. Before the start of the experiment, the participants were provided with an informed consent form. Upon consenting, a brief description of the procedure was provided. Then, the eligibility of the participants was re-confirmed through a questionnaire. Then, the participants answered a questionnaire assessing demographic variables, their living conditions, the effect of COVID-19 on their daily life, VUI usage experience[2], prior self-compassion and loneliness scores.

After that, the control condition participants answered the questionnaire corresponding to stage 1 of the control condition. This concluded the first stage of the control condition. Meanwhile, the participants in the CR and CG condition were redirected to a webpage where VA was hosted. Here, they were provided with elaborate instructions on how to interact with VA. Upon submitting their participant ID, VA was enabled for conversation. After the conversation, the participants were redirected to another questionnaire that assessed the participants' engagement with VA. This questionnaire also inquired if the participants faced technical problems. If the participants chose to drop out, they were directed to the same questionnaire displaying the NHS weblink and hotline number. This questionnaire contained a conditional section that appeared when the

[2] Only for CR and CG condition

participant chose to drop out, inquiring if they are comfortable sharing their reasons for dropping out. This section of the questionnaire was made optional. This concluded the first stage for both CR and CG conditions.

After five working days[3], an invite was sent to the same participants for the second stage. For the control condition, the participants answered a questionnaire. This concluded the second stage of the control condition. Meanwhile, stage 2 of CG and CR conditions were conducted one after another. In the CR and CG conditions, the participants engaged again in a conversation with VA, followed by a questionnaire assessing their engagement. Like the previous stage, this questionnaire also inquired if the participants dropped out or faced any technical problems. This concluded the second stage for both CR and CG conditions.

After three working days, an invite was sent to the same participants for the final stage. For the control condition, the participants answered a questionnaire. This questionnaire also assessed the self-compassion and loneliness score at the end of the experiment. This concluded the final stage for the control condition. In the CR and CG conditions, the participants engaged in a conversation with VA, followed by a questionnaire assessing their engagement. Like the previous stage, this questionnaire also asked whether the participants faced any technical difficulties or if they dropped out of the experiment. Further, this questionnaire also measured their self-compassion and loneliness scores at the end of the experiment, their perception of VA, and the conversation. The questionnaire also inquired if COVID-19 had significantly affected their daily life over the course of the experiment. With this, the final stage was concluded. Finally, participants were debriefed, compensated and thanked for their participation.

---

[3] Initially, the first stage of all the conditions were conducted simultaneously. During the first stage, technical problems were encountered concerning VA being unable to handle multiple people talking simultaneously. While troubleshooting this problem, the experiment was put on hold for a week. After that, each stage was conducted for three days. Furthermore, a stage of only one of CR or CG conditions was conducted at a time due to limitations on computational power.

Pre-processing

Engagement Proxy. To assess the engagement of the participants in the study, an engagement proxy was calculated based on the qualitative data[4]. The proxy was calculated in the following manner: (1) For each stage, the number of characters in the qualitative data were measured per condition, (2) for each participant, the average number of characters called engagement score across the stage was computed, finally, (3) for each condition, every participant was assessed if their engagement score was above the first quartile. Participants with an engagement score above the first quartile were assigned 'high engagement', and the rest were assigned 'low engagement'. This means that participants with engagement scores higher than the bottom 25% of the engagement scores were considered highly engaged in their interactions with VA.

Data Analysis

Data analysis was done in STATA IC 16.0 and RStudio 1.2.5033 (R version 3.6.3) (R Core Team, 2019). Since the experiment design involves assessing the relative effectiveness of 3 conditions with 2 points of measurement, mixed ANOVA was chosen to be the appropriate statistical test. In case of a significant interaction effect, post-hoc contrasts were analyzed.

Non-parametric analysis. In case of violation of assumptions about normality and homogeneity of variances in the construct of interest, i.e. self-compassion and loneliness, non-parametric rank means tests were conducted using **'nparLD'** package (Noguchi, Gel, Brunner, & Konietschke, 2012) in R studio. The nparLD package generates the output for a non-parametric mixed model analysis in terms of Relative Treatment Effects (RTE) and ANOVA Type Statistics (ATS)(Bathke, Schabenberger, Tobias, & Madden, 2009). The RTEs can be interpreted in the following manner: For example, the RTE for CG condition is 0.56, which means a randomly chosen observation from the

[4] Only 20% of the open-ended questions in the conversations with VA were transcribed successfully. The speech to text algorithm's accuracy was compromised due to noise and other uncontrollable factors as the study was not conducted in a laboratory.

whole dataset results in a smaller value than a randomly chosen observation from the CG condition with an estimated probability of 56%.

Equivalence Testing. If no interaction effect is found to be significant, equivalence testing was done by means of Two One-Sided T-tests (TOST) procedure (Lakens, 2017). Often, the absence of a significant result in the null hypothesis significance testing is incorrectly reported as the absence of an effect. It is to be noted that the null hypothesis significance testing allows only for the rejection of a null hypothesis and not its support. This implies that we cannot claim for an effect to be zero. On the other hand, equivalence testing allows for testing if the effect size falls within a specified range of effect sizes that are close to zero and that any value within these bounds can be statistically regarded as being equivalent to zero (Lakens, 2017). Equivalence tests are conducted by setting equivalence bounds. The equivalence bounds were set based on the SESOI($f(U) = 0.19$) and available resources. For this study, the bounds were set at d_z = -0.40 to d_z = 0.40 ($f(U) = -0.19$ to $f(U) = 0.19$).

Qualitative analysis. The qualitative data from the study were analysed using thematic analysis (Braun & Clarke, 2012). After familiarising with the data, the data were coded to describe the idea or feeling expressed in those parts. Then, themes were generated based on the codes. After that, the themes were reviewed to ensure the fit of the codes with the themes. Finally, the themes were named and elaborated with supporting participant quotes.

Ethical Considerations

In cases that the participants expressed discomfort while participating in the study, they were provided with a weblink and a hotline number to access the National Health Service (NHS) provided by the UK government. The ethical board of the Human-Technology Interaction department at the TU/e approved of this method of implementation for the current study.

4. Implementation

This section provides information regarding the development and deployment of VA. First, the details about the technical components will be discussed. Second, the process flow between constituent components of VA will be discussed. Third, the deployment of the webpage will be discussed. Fourth, challenges and suggestions for future technical development will be discussed. Finally, the details corresponding to the pilot test that helped decide the gender-ambiguous voice of VA will be discussed.

Technical components

Front-end. The front-end, also known as the visual component of a webpage, was developed using the **Angular framework** (Angular team, 2016). It is an open-source application design framework and development platform that uses TypeScript for creating single-page web apps. The front-end was designed to display only VA's audio responses to avoid hurting the participants' perception of VA's human-like persona due to inaccurate transcription. It is to be noted that the participants' responses were transcribed only for the researcher's purposes and were not displayed in VA's chat window.

Back-end. The back-end, also known as the serve-side of the VUI, is the processing component. The back-end is made of two units: (1) the back-end of the webpage developed using Python, and (2) Google Dialogflow API (Google, 2016). These two units communicate back and forth along with the front end to keep the VUI processing the participants' dialogue and responding accordingly. The webpage's back-end was used to transfer the audio between the front-end and the Google Dialogflow API. **Google Dialogflow** is a natural language understanding platform used to develop conversational user interfaces(CUIs). Dialogflow converts the audio spoken by the user to text, processes it to find the intent, and replies with an online action or reply in the form of text or audio. The conversation design for VA was done using the Dialogflow console. VA was developed as a Dialogflow ES agent[5]. The agent developed on Dialogflow was enabled to work on the webpage as third-party integration.

[5] Refer to the Google Dialogflow documentation

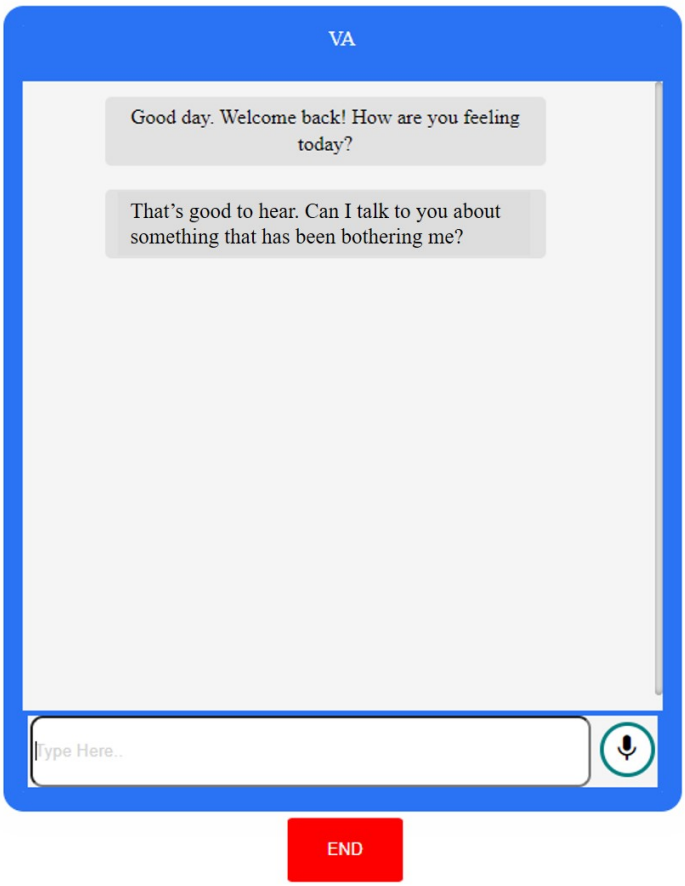

Figure 2. Screenshot of VA on the webpage

To protect the participants' privacy, the conversational agent that runs VA was designed and hosted in one of Dialogflow regional data centres in the UK[6].

Process flow

When the participant converses with VA by speaking or typing an expression, the front end relays the text or audio to the webpage's back-end using Flask[7](Armin Ronacher, 2010) and Flask-Socket IO[8](Miguel Grinberg, 2018) integration. The audio

[6] Dialogflow provides data residency to keep user data-at-rest physically within a geographical region.

[7] Flask is a microweb framework written in Python.

[8] Flask-Socket.IO is a Socket.IO integration for Flask applications. Socket.IO is a protocol that enables ensures a low latency bi-directional communication between the clients and servers.

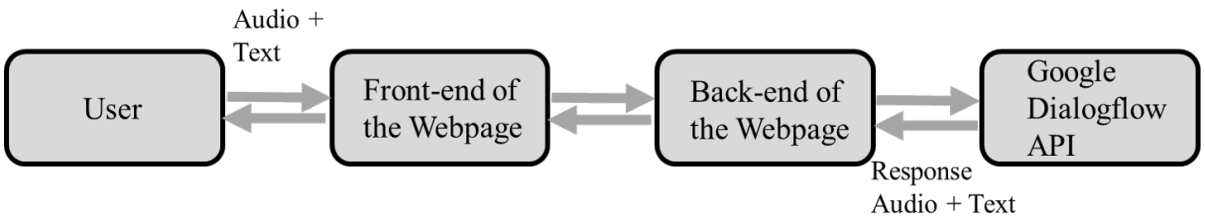

Figure 3. Process flow between the technical components

is then sent as an audio stream using a third-party integration code[9] in Python to Dialogflow in a detect intent request message. Dialogflow sends a detect intent response message to the webpage back-end. This message contains information about the matched intent, the action, the parameters, and the response defined for the intent. The back-end relays this response to the front-end, and the participant sees or hears the response.

Challenges and alternative solutions

First, the challenges that were encountered will be discussed. Second, solutions that were implemented to address the challenges will be discussed. Finally, alternative methodologies or software components to make the process flow robust will be discussed.

The first challenge encountered with the current process flow entailed momentary discontinuation of the conversation when multiple participants tried speaking simultaneously. This was most likely because Flask handles data in a serial manner, i.e., one packet of data after another. Therefore, when multiple participants tried to speak with VA simultaneously, only a few of the participant-side audio responses could be processed. The second challenge was that VA was limited to handling a simple linear flow of conversation[10]. The third challenge was that the time available per participant for each dialogue with VA was limited to 25 seconds. This is a limit set by the Dialogflow API when a VUI is setup as a third-party integration.

---

[9] Refer to API Interaction in Google Dialogflow documentation at
https://cloud.google.com/dialogflow/es/docs/api-overview

[10] At the time of development of VA, only ES type of agents could be developed. For more information, refer to Google Dialogflow documentation

To address the first challenge, a brute force approach was used. Multiple webpage back-ends were initialized for one front-end to enable around 10 participants to talk to VA simultaneously. The downside with this approach was that the Virtual Private Server's (VPS) processing power could not handle the load of two VAs running at the same time for two different conditions. Many intents corresponding to various possible participant responses were created to address the second challenge, making it a high-order decision tree, and yet, conversation repair seemed problematic. The third challenge can be addressed only by purchasing a premium edition of Dialogflow.

Alternatively, to address the first challenge, one could use Gunicorn (Benoit Chesneau, 2010). This approach was attempted, but it did not yield any success due to Flask-Socket.IO and Gunicorn integration problems. An alternative approach to address the second challenge would be to develop a practical conversation repair framework and also use a finite-state machine model[11] (Bors, 2018) for navigating through the conversation. At the time of VA's development, there were no provisions for developing a VUI with state machine logic. It is now possible by using the Dialogflow CX Agent configuration, released on December 15, 2020.

Deployment

The webpage with VA was hosted on a VPS setup on a data centre located in the Netherlands to ensure GDPR compliance. The VPS was running Ubuntu 18.04 64 bit operating system with 4GB of RAM.

Pilot test - Voice

To decide the gender-ambiguous voice of VA, a pilot test was conducted. Fourteen participants - seven men, six women, and one gender undisclosed participants took part in the pilot study. They listened to 9 audio clips generated by altering the pitch of the available synthetic voices provided by Dialogflow. The participants answered a questionnaire

[11] A Finite State Machine is a model of computation based on a hypothetical machine made of one or more states. Only one single state of this machine can be active at the same time.

indicating the perceived gender of voice in the audio clips as male, female, or indeterminable. The audio clip that most participants (57%) answered to be indeterminable was chosen as VA's voice.

5. Results

Descriptives

Table 3

VUI usage experience

	CG		CR		All	
	M	SD	M	SD	M	SD
VUI usage experience	3.21	1.13	3.06	1.12	3.15	1.12

The median of participants' experience with VUI usage was around 3 on a 5-point frequency scale (ranging from 'Never to 'Daily') for the CR and CG conditions as summarized in Table 3. Before the start of the experiment, the average self-compassion score was 3.92 (SD=0.83) for all conditions. The average prior loneliness score was measured to be 21.28 (SD=7.04). Table 4 shows the average self-compassion and loneliness scores in each condition before the start of the experiment.

The observed prior self-compassion scores rejected normality in the control condition. Therefore, a Kruskal Wallis test was conducted to test if there were significant differences in self-compassion across conditions before the experiment. None of the differences were statistically significant. A one-way ANOVAs tested whether there were significant differences in loneliness scores across conditions prior to the experiment. No statistically significant differences were found.

Table 4

Prior self-compassion and loneliness scores

	CG		CR		Control		All	
	M	SD	M	SD	M	SD	M	SD
Prior self-compassion	3.91	.82	4.07	0.92	3.86	.80	3.92	.83
Prior loneliness	21.67	7.56	21.32	7.28	21.09	6.79	21.28	7.04

Perception of VA and Conversation. Overall, VA was perceived to be moderately caring, likeable, trustworthy and intelligent across all conditions. The average

scores on these 10-point scales were higher than 6 but on no scale did VA get an average score of more than 7. There is not much variation in the average scores on the following five perception variables: (1) Genderedness, (2) Caring, (3) Likability, (4) Trustworthiness, and (5) Intelligence. Intuitively, the average score for dominance was lower in the CR condition compared to the CG condition. Also, the average submissiveness score in the CR condition was greater than the CG condition. Since VA provided care in the CG condition, it can be expected to be perceived as more dominating. However, in the CR condition, VA asks for care which could make it appear more submissive. Regarding the perception of conversation, VA was perceived to be less responsive and only slightly evoked the perception of having a real conversation. Also, VA was perceived to be different from other VUIs.

Five two-tailed independent t-tests across the variables - genderedness, caring, likability, trustworthy, and intelligence were conducted. Two one-tailed t-tests were conducted with the expectation that VA is more dominant in CG condition and more submissive in CR condition. Using a Bonferroni corrected alpha level of 0.05/7=0.007, the five two-tailed t-tests yielded no significant results. As expected, the 31 participants in the CR condition perceived VA to be significantly more submissive compared to the 42 participants in the CG condition, $t(71) = 2.54$ ($p<.007$). The one-tailed t-test for dominance did not yield any significant results. Further, Fischer's exact test showed a significant association ($p<.05$) between the condition and the perceived gender of VA. From Table 6, it can be inferred that the participants in the CR condition perceived VA as male, and the participants in the CG condition perceived VA as female.

Table 6

Perceived gender of VA

	CG	CR
Male	14	20
Female	15	8
Other	13	3

Table 5

Perception of VA and Conversation

	H					
	CG		CR		All	
	M	SD	M	SD	M	SD
Genderedness	5.61	3.05	5.93	2.72	5.75	2.90
Caring	6.70	1.45	6.73	1.77	6.72	1.58
Likability	6.58	1.66	6.12	2.32	6.38	1.92
Trustworthiness	6.55	1.65	6.52	1.83	6.53	1.71
Intelligence	6.44	1.84	6.21	1.99	6.34	1.89
Dominance	4.92	1.62	4.05	1.83	4.55	1.75
Submissiveness	5.12	1.41	6.07	1.60	5.56	1.55
VA listened and responded to what I said	4.71	1.47	3.97	1.83	4.40	1.67
I felt I was having a real conversation	3.17	1.59	2.38	1.47	2.83	1.58
VA's responses resembled those of other bots	4.59	1.38	2.39	1.47	2.83	1.58

Outliers

Data points pertaining to the dependent variable that lie outside 1.5 times the interquartile range above the upper quartile and below the lower quartile were identified as outliers (Barbato, Barini, Genta, & Levi, 2011). Further, for variables with normal distribution, data points whose standardized values of the dependent variable were greater than 3 and lesser than -3 were also considered outliers. Based on these criteria, 11 outliers were identified when analyzing for self-compassion and, 3 outliers were identified when analyzing for loneliness. The remainder of the analysis was done excluding these outliers.

Hypotheses testing

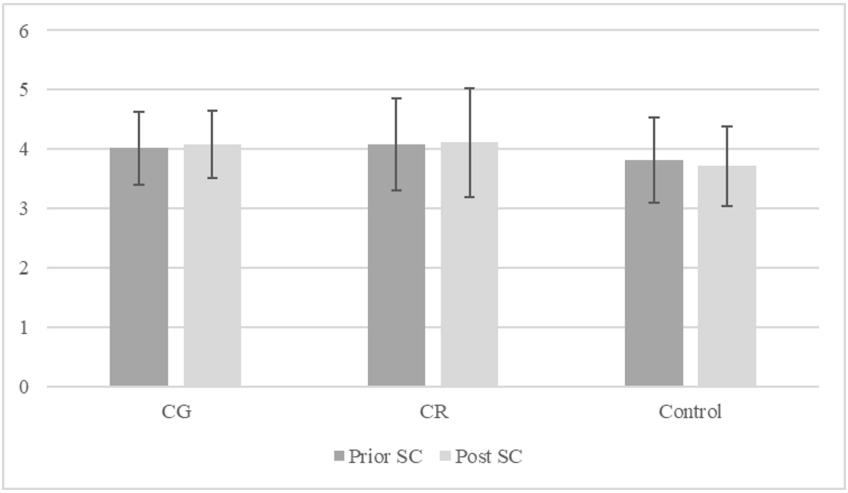

Figure 4. Means of Prior and Post Self-Compassion scores across conditions

Table 7

Means(Standard Deviation) of observed self-compassion scores

	CG	CR	Control	All
Prior self-compassion	4.02(0.61)	4.08(0.77)	3.81(0.72)	3.91(0.71)
Post self-compassion	4.08(0.57)	4.10(0.92)	3.71(0.68)	3.88(0.73)

Self-compassion. The observed post self-compassion scores reject the assumption of homogeneity of variances (Levene's test at the mean, $F(2,147)=3.31$, $p < 0.05$). Since the assumption of homogeneity of variances cannot be upheld, a non-parametric rank means test for mixed designs was conducted using the 'nparLD' (Noguchi et al., 2012) package in R. Self-compassion scores were expected to increase with time across all conditions, with the CR condition showing the greatest improvement in self-compassion and the control condition showing the least. The main effect of condition on self-compassion (ATS[12](1.47)=2.98, p=.058) was not statistically significant. Further, there was no a statistically significant main effect of time on self-compassion ($ATS(1)$=0.41, p=.52), nor a statistically significant interaction effect of time and condition ($ATS(1.96)$=1.57, p=.21)

[12] ANOVA Type Statistic (ATS)

on self-compassion[13].

Further, the change in self-compassion score (the difference between post self-compassion score and the prior self-compassion score) was calculated for each participant to compensate for individual differences. Table 8 shows the averages of the change in self-compassion scores across conditions. It was expected that there would be significant differences in the change in self-compassion scores across conditions, with the CR condition showing the greatest change in self-compassion and the control condition showing the least change in self-compassion. A one-way ANOVA was conducted to test if there were any significant differences in change in self-compassion score across conditions. No statistically significant effects were found[14].

Table 8

Means(Standard Deviation) of change in self-compassion scores across conditions

	CG	CR	Control	All
Change in self-compassion	0.028(0.54)	0.052(0.52)	-0.12(0.54)	-0.048(0.537)

Table 9

ANOVA: output for change in self-compassion score

	Type III sum of squares	df	Mean square	F	Sig.
Condition	1.01	2	0.51	1.78	.1725
Residual	44.62	156	0.29		

[13] No significant effects were found with the outliers included

[14] No significant effects were found with the outliers included

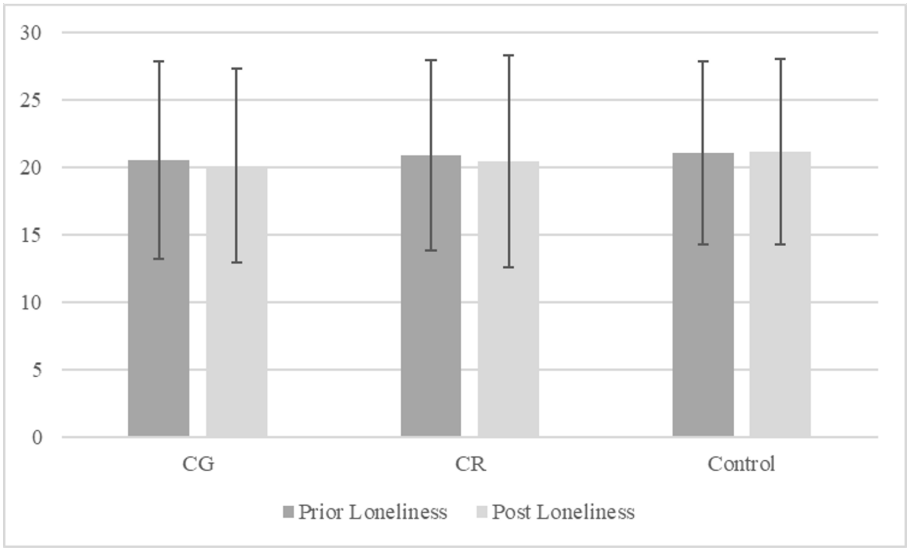

Figure 5. Prior and Post Loneliness scores across conditions

Loneliness. Table 10 shows the average values of observed prior and post loneliness scores. Loneliness scores were expected to decrease with time across all conditions, with the CR condition showing the greatest decrease in loneliness and the control condition showing the least. The observed post loneliness scores in the CR condition reject the assumption of normality (W=0.93). The non-parametric rank means test using 'nparLD' package was conducted. There was no statistically significant main effect of condition ($ATS(1.76)$=0.034, p=.95), time ($ATS(1)$=4, p=.31), nor a statistically significant interaction effect of time and condition ($ATS(1.88)$=0.99, p=.366) on loneliness scores[15].

Table 10

Means (Standard Deviation) of observed loneliness scores

	CG	CR	Control	All
Prior loneliness score	21.34(7.35)	21.32(7.27)	20.72(6.40)	21(6.79)
Post loneliness score	20.19(6.92)	21.35(8.2)	20.74(6.37)	20.72(6.87)

Further, the change in loneliness score (the difference between post loneliness score and the prior loneliness score) was calculated for each participant to compensate for large individual differences. Table 11 shows the averages of the change in loneliness scores

[15] No significant effects were found with the outliers included

across conditions. It was expected that there would be significant differences in the change in loneliness scores across conditions, with the CR condition showing the greatest change in loneliness and the control condition showing the least change in loneliness. A one-way ANOVA was conducted to test if there were any significant differences in change in loneliness score across conditions. No statistically significant effects were found[16].

Table 11

Means(Standard Deviation) of change in loneliness scores across conditions

	CG	CR	Control	All
Change in loneliness	-1.05(3.80)	-0.4(3.62)	0.068(4.09)	-0.31(3.93)

Table 12

ANOVA: output for change in loneliness score

	Type III sum of squares	df	Mean square	F	Sig.
Condition	35.68	2	17.84	1.15	.318
Residual	2428.70	157	15.47		

Equivalence testing

Self-Compassion. The main effect of condition and time were not statistically significant. Three equivalence Two One-Sided T-test (TOST) procedures were conducted to determine if the effect sizes of change in self-compassion scores within a condition are equivalent to zero. The equivalence bounds were set at d_z=-0.4 and d_z=0.4,(or f=-0.21 and f=0.21). For the CG condition, the TOST procedure indicated that the observed effect size (d_z=0.12) was significantly within the equivalent bounds ($t(36)$=1.72, p=.004). For the CR condition, the TOST procedure indicated that the observed effect size (d_z=0.05) was significantly within the equivalent bounds ($t(28)$=-1.91, p=.034). For the control condition, the TOST procedure indicated that the observed effect size (d_z=-0.19) was significantly within the equivalent bounds ($t(28)$=1.91, p=.003). The TOST procedures indicated that the observed effect sizes of change in self-compassion within all

[16] No significant effects were found with the outliers included

the conditions were significantly within the equivalence bounds, meaning that the effect sizes were equivalent to zero.

The interaction effect of time and condition was not statistically significant. A second set of three TOST procedures were conducted to determine if the effect sizes corresponding to relative differences between the conditions was indeed absent (Lakens, 2017). Table 13 shows the average change in self-compassion scores (calculated as a difference between post and prior self-compassion scores) and the effect sizes corresponding to the change in self-compassion scores between conditions. The equivalence bounds were set at d_z=-0.4 and d_z=0.4,(or f=-0.21 and f=0.21). The first TOST procedure based on Student's t-test indicated that the observed effect size of difference (d_z=-0.07) between CR and CG conditions was not significantly equivalent ($t(64)$=1.31, p=.097) to zero within the equivalence bounds. Between control and CG conditions, the second TOST procedure based on Welche's t-test[17] indicated that the observed effect size of difference(d_z=-0.32) was not significantly equivalent ($t(119)$=0.40, p=.343) to zero within the equivalence bounds. The third TOST procedure based on Welche's t-test indicated that the observed effect size of difference effect size of difference(d_z=-0.25) between control and CR conditions was not significantly equivalent ($t(111)$=0.72, p=.237) to zero within the equivalence bounds. Though these effect sizes are not equivalent to zero, the effect size of self-compassion within every condition is zero indicating that there are no relative differences.

[17] Though the change in self-compassion score was distributed normally across conditions, the disparity in the sample size warrants a non-parametric test

Table 13

Changes in self-compassion per condition

Condition	Changes in self-compassion			Effect size d_z		
	n	M	SD	CG	CR	Control
Care-Giving(CG)	37	0.064	0.51	x		
Care-Receiving(CR)	29	0.026	0.51	-0.07	x	
Control	84	-0.10	0.52	-0.32	-0.25	x

Exploratory analyses

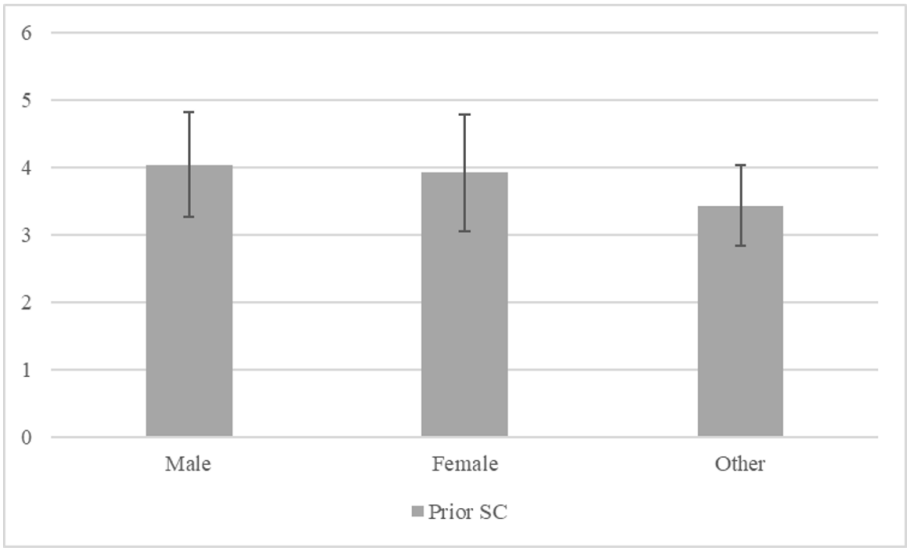

Figure 6. Prior Self-Compassion scores across genders

Prior self-compassion scores between male and female paricipants. Table 14 shows the average self-compassion scores for male and female participants across conditions. Prior self-compassion scores are expected to be lower in female participants compared to male participants (Yarnell et al., 2015). The observed prior self-compassion scores follow a normal distribution for both male and female participants. Since the number of female participants is twice as large as the number of male participants, Mann-Whitney U test was conducted. There is no significant difference between male and female participants with regard to the observed prior self-compassion scores (z=1.47, p=.14).

Table 14

Means(Stand Deviation) of self-compassion scores for male and female participants across conditions

Male participants	CG	CR	Control	All
Prior self-compassion	4.14(0.60)	4.63(0.64)	3.74(0.73)	4.04(0.75)
Post self-compassion	4.10(0.91)	4.77(1.02)	3.72(0.81)	4.05(0.96)
Change in self-compassion	0.10(0.60)	0.14(0.60)	-0.02(0.60)	0.492(0.60)
Female participants	CG	CR	Control	All
Prior self-compassion	3.96(0.60)	3.77(0.90)	3.81(0.67)	3.83(0.70)
Post self-compassion	3.98(0.53)	3.76(0.71)	3.68(0.51)	3.76(0.57)
Change in self-compassion	-0.09(0.45)	-0.05(0.44)	-0.14(0.48)	-0.11(0.46)

Self-compassion in males. Table 14 shows the average values of prior, post and change in self-compassion scores in male participants. The assumption for normality was rejected by the observed prior self-compassion scores in CR and control conditions. Non-parametric rank means test using 'nparLD' package was conducted. The main effect of condition on self-compassion in males was statistically significant($ATS(1.97)$=13.06, p<.000). There was no statistically significant main effect of time on self-compassion ($ATS(1)$=0.53, p=.463), nor a statistically significant interaction effect of time and condition ($ATS(1.99)$=0.35, p=.703) on self-compassion. It is important to note that this analysis is based on medians and not means of the observed self-compassion scores as the analysis is non-parametric. The RTEs concerning each condition, time point, and condition - time interaction, are shown in Table 15.

Point estimates of Relative Treatment Effects, also known as Relative Treatment Effects (RTEs), can be interpreted in the following manner: For example, the RTE for CG condition is 0.50, which means a randomly chosen observation from the whole dataset results in a smaller value than a randomly chosen observation from the CG condition with an estimated probability of 50%.

The change in self-compassion scores (the difference between post self-compassion

score and the prior self-compassion score) was calculated for each male participant to compensate for individual differences. Table 14 shows the averages of the change in self-compassion scores in male participants across conditions. A one-way ANOVA was conducted to test if there were any significant differences in change in self-compassion score across conditions. No statistically significant effects were found.

Table 15

RTEs of the non-parametric model on self compassion in males

	CG	CR	Control	time1	time2
RTE	0.50	0.77	0.40	0.54	0.57

	CG:time1	CG:time2	CR:time1	CR:time2	Control:time1	Control:time2
RTE	0.48	0.53	0.75	0.78	0.40	0.40

Post-hoc analyses:

The assumption of normality was rejected in the observed prior self-compassion scores of male participants in the CR and the control conditions. A Kruskal-Wallis test was conducted and the prior self-compassion scores in male participants showed statistically significant difference across conditions ($\chi^2(2,51)=10.80$, $p=.004$). Further, pair-wise comparison using Mann-Whitney U test revealed that the post self-compassion scores of male participants in the CG condition were statistically significantly lower than the post self-compassion scores of male participants in the CR condition ($Z(23)=2.73$, $p=.005$). Post self-compassion scores in male participants showed statistically significant difference across conditions ($F(2,48)=9.13$, $p<.000$). Further, pair-wise comparison revealed that the post self-compassion scores of male participants in the control condition were statistically significantly lower than the post self-compassion scores of male participants in the CR condition ($t(37)=4.23$, $p<.000$).

Self-compassion in females. Table 14 shows the average values of prior, post and change in self-compassion scores in female participants. Since the assumptions of normality and homogeneity of variances were not rejected, mixed ANOVA was conducted. No statistically significant effects were observed.

The change in self-compassion scores of female participants (the difference between post self-compassion score and the prior self-compassion score) was calculated for each participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in self-compassion score across conditions. No statistically significant effects were found.

Relation between VA's perceived gender and gender of the participants. In the CG condition, nearly half (43%) of the male participants perceived VA to be male and nearly half (41%) of the female participants perceived VA to be female. In the CR condition, most (83%) of the male participants perceived VA to be male and half of the female participants perceived the same. Table 16 shows the frequency table indicating the relation between the gender of the participants and the perceived gender of VA. No statistically significant associations were found between perceived gender of VA and gender of the participants by conducting Fischer's exact tests in the CG(p=.360) and CR(p=0.339) conditions. While there are no statistically significant associations, it is interesting to note the ambiguity in VA's perceived gender by female participants in the CG condition.

Table 16

Perceived gender of VA by gender of the participants in CG and CR conditions

CG	VA as male	VA as female	Can't say
Male participants	7	5	4
Female participants	7	10	7
Other participants	0	0	2
CR	VA as male	VA as female	Can't say
Male participants	10	2	0
Female participants	9	6	3
Other participants	1	0	0

Table 17

Means(Stand Deviation) of loneliness scores between genders across conditions

Male participants	CG	CR	Control	All
Prior loneliness	20.6(6.59)	19.67(5.90)	19.54(5.99)	19.87(6.04)
Post loneliness	19.33(4.40)	18.50(7.67)	19.38(6.30)	19.17(6.06)
Change in loneliness	-1.56(3.55)	-1.16(3.71)	-1.78(3.84)	-0.78(3.72)
Female participants	CG	CR	Control	All
Prior loneliness	22.67(8.45)	21.83(7.87)	20.93(6.18)	21.50(7.04)
Post loneliness	21.87(8.73)	22.61(8.08)	21.08(6.12)	21.54(7.12)
Change in loneliness	-0.80(4.13)	0.06(3.68)	0.15(4.25)	-0.09(4.11)

Loneliness in males. Table 17 shows the average values of observed prior, post and change in loneliness scores in male participants across conditions. The change in loneliness score (the difference between post loneliness score and the prior loneliness score) was calculated for each male participant to compensate for large individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in loneliness scores across conditions. No statistically significant effects were found.

Loneliness in females. Table 17 shows the average values of observed prior, post and change in loneliness scores in female participants across conditions. The change in loneliness score (the difference between post loneliness score and the prior loneliness score) was calculated for each female participant to compensate for large individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in loneliness scores across conditions. No statistically significant effects were found.

Common Humanity. Stage 1 of all the conditions involved tasks to stimulate self-compassion along the dimension of common humanity. Accordingly, it was expected that the scores of common humanity sub-scale would increase across all conditions, with the participants in the CR condition showing the greatest improvement in common humanity scores, and the control condition the least. Table 18 shows the average values

of observed prior, post and change in common humanity scores across conditions. The change in common humanity scores (the difference between post common humanity score and the prior prior humanity score) was calculated for each participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in common humanity score across conditions. No statistically significant effects were found.

Table 18

Means(Stand Deviation) of common humanity scores across conditions

All participants	CG	CR	Control	All
Prior common humanity	4.76(1.02)	5.06(0.88)	4.38(1.03)	4.60(1.03)
Post common humanity	4.70(0.92)	4.94(0.78)	4.16(0.97)	4.44(0.98)
Change in common humanity	-0.11(0.84)	-0.12(0.82)	-0.12(0.88)	-0.11(0.86)
Male participants	CG	CR	Control	All
Prior common humanity	5.04(0.43)	5.29(0.98)	4.37(1.15)	4.72(1.06)
Post common humanity	5.10(0.64)	5.16(0.75)	4.04(1.20)	4.53(1.13)
Change in common humanity	-0.06(0.80)	-0.25(0.95)	-0.21(1.12)	-0.18(0.99)
Female participants	CG	CR	Control	All
Prior common humanity	4.50(1.27)	4.97(0.80)	4.43(1.00)	4.54(1.05)
Post common humanity	4.43(1.07)	4.85(0.78)	4.33(0.97)	4.44(0.97)
Change in common humanity	-0.07(1.00)	-0.10(0.52)	-0.10(0.84)	-0.09(0.84)

Common Humanity in males. Table 18 shows the average values of observed prior, post and change in common humanity scores in male participants across conditions. The change in common humanity scores (the difference between post self-compassion score and the prior self-compassion score) was calculated for each male participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in common humanity score across conditions. No statistically significant effects were found.

Common Humanity in females. Table 18 shows the average values of observed prior, post and change in common humanity scores in female participants across conditions. The change in common humanity scores (the difference between post self-compassion score and the prior self-compassion score) was calculated for each female participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in common humanity score across conditions. No statistically significant effects were found.

Table 19

Means(Stand Deviation) of mindfulness scores across conditions

All participants	CG	CR	Control	All
Prior mindfulness	4.76(0.82)	4.87(0.82)	4.42(0.96)	4.59(0.92)
Post mindfulness	4.67(0.89)	4.75(0.99)	4.26(0.98)	4.45(0.98)
Change in mindfulness	-0.21(0.73)	0.04(0.80)	-0.18(0.72)	-0.15(0.74)
Male participants	CG	CR	Control	All
Prior mindfulness	5.00(0.66)	5.27(0.29)	4.54(0.78)	4.80(0.74)
Post mindfulness	5.00(0.62)	5.36(1.00)	4.32(0.83)	4.70(0.90)
Change in mindfulness	0.00(0.80)	0.09(0.78)	-0.22(0.72)	-0.10(0.75)
Female participants	CG	CR	Control	All
Prior mindfulness	4.67(0.90)	4.50(0.75)	4.38(1.04)	4.46(0.97)
Post mindfulness	4.49(1.02)	4.41(0.83)	4.24(1.06)	4.32(1.00)
Change in mindfulness	-0.23(0.90)	0.03(0.85)	-0.14(0.80)	-0.13(0.83)

Mindfulness. Stage 3 in all the conditions involved tasks to stimulate self-compassion along the dimension of mindfulness. Table 19 shows the average values of observed prior, post and change in mindfulness scores across conditions. The change in mindfulness scores (the difference between post mindfulness score and the prior mindfulness score) was calculated for each participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in mindfulness scores across conditions. No statistically significant effects were found.

Mindfulness in males. Table 19 shows the average values of observed prior, post and change in mindfulness scores in male participants across conditions. The change in mindfulness scores of male participants (the difference between post mindfulness score and the prior mindfulness score) was calculated for each male participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in mindfulness scores across conditions. No statistically significant effects were found.

Mindfulness in females. Table 19 shows the average values of observed prior, post and change in mindfulness scores in female participants across conditions. The change in mindfulness scores of female participants (the difference between post mindfulness score and the prior mindfulness score) was calculated for each female participant to compensate for individual differences. A one-way ANOVA was conducted to test if there were any significant differences in change in mindfulness scores across conditions. No statistically significant effects were found.

Engagement. Engagement with VA was measured through the calculation of engagement proxy. Table 20 shows the number of male and female participants per condition based on their level of engagement. Fischer's exact tests revealed that there was no statistically significant association between gender of participants and condition both in high(p=.875) and low(p=.846) engagement groups.

Table 20

Engagement of participants between conditions based on gender

High Engagement	CG	CR
Male Participants	10	8
Female Participants	21	15
Other pariticipants	0	0
Low Engagement	CG	CR
Male Participants	6	4
Female Participants	3	3
Other pariticipants	2	1

Exploratory Qualitative Analysis

To understand more about the nature of conversation between VA and the participants, the qualitative data was analysed using thematic analysis (Braun & Clarke, 2012). Our corpus consisted of: (1) Transcribed dialogues between VA and the participant, (2) Answer to the open-ended question about the perception of conversation with VA at the end of stage 3 of the study: "Why did VA's responses (not) resemble other voice-bots/assistants?", and (3) Response to the measure Effect of COVID-19, an open-ended question asked at the beginning and the end of the study to account for any significant changes in lives of the participants over the course of the study: "How has the current COVID 19 pandemic affected you in the duration of this study (since you started interacting with VA)?".

This analysis is to be considered exploratory due to discrepancies[18] in transcription. Further, the amount of time available to a participant for each dialogue with VA was limited to 25 seconds per dialogue. The Dialogflow API places this time limit on VUIs that use third-party integration. As an effect, the qualitative data available per participant in the CR and CG conditions is less compared to the qualitative data from the control condition.

Upon analysing the qualitative data, five main themes come to light. First, human-like or robotic perception of VA. Second, the emotiveness of the voice. Third, the lack of judgement. Fourth, misunderstanding and unresponsiveness. Finally, the specificity of responses.

Human-like versus Robotic. In stage 1, the CG condition participants talked to VA about a personal moment of rejection or humiliation. In one of the steps of the task, they were asked to imagine a friend undergoing a similar situation and offer words of compassion. In stage 1 of CR condition, VA asked the participants for some advice on a situation it was having difficulty dealing with. Interestingly, the responses in both

[18] Only 20% of the open-ended questions in the conversations with VA were transcribed successfully. The accuracy of the speech to text algorithm was compromised due to noise and other uncontrollable factors as the study was conducted outside a laboratory.

conditions were similar. For example, a participant in the CG condition offered the following advice to an imaginary friend:

"...I would tell them its okay, everything will be fine. don't worry about it."

In the CR condition, a participant advised VA in the following manner:

"Stay positive. The situation should improve once COVID is tackled."

Some participants spoke to VA perceiving it as human or human-like. For example, the following phrases were used to advise VA:

"Try to find enjoyment in the small things in life... ", "I think you need to see the positives in life."

On the other hand, some participants refused to offer VA any advice perceiving it as a robot. For example, a participant who was unwilling to offer advice to VA said the following:

"...you are a robot you don't have feelings".

Based on the responses to the open-ended question: "Why did VA's responses (not) resemble other voice-bots/assistants?", we find opposing views on the perceived nature of VA's voice. While some participants expressed that VA's voice sounded natural and human-like, some participants felt that VA's voice was robotic and computerised. For example, participants in the CG condition shared the following when asked about VA's responsiveness:

"it sounded human-like...", "language was very human"

Another participant in the same condition responded in the following manner:

"Very robotic and formal, felt very structured."

Emotive vs Lifeless . Not all participants enjoyed their conversation with VA. Some participants perceived VA as compassionate, emotion-filled, warm and friendly compared to other VUIs. For example, considering the perceived emotional intelligence of VA, a participant shared the following opinion:

"It seemed more compassionate and understanding."

In the CR condition, a participant expressed their views on VA's voice in the following manner:

"The voice was a bit whiny, as if it was very timid and sensitive."

In contrast, some participants perceived VA as lifeless, weird and scripted. Consider the following responses from a few participants:

"They were lifeless and misunderstanding of nuance."

"They were definitely pre-recorded..."

Lack of judgement. Though participants' perception of VA varied from being smart, compassionate, and human-like to scripted, lifeless, and weird, the participants showed no reservations in disclosing their thoughts and feelings. Consider the sensitive nature of the content shared by participants in the CG condition:

"My husband isn't supporting me with my daughter. I am sad and trapped, I would like to leave because the lack of support isn't going away. I feel diminished."

"I was okay with my decision to leave my job because it wasn't right for me but I was feeling stressed because of the money."

Misunderstanding and unresponsiveness. Some participants did not have a pleasant experience interacting with VA. This was either due to VA misunderstanding the participants' responses or failure of conversation repair. The following participants' responses shed light on their experiences interacting with VA:

"It was trying to understand but couldn't and was not personalised."

"..there were scripted answers that almost but not totally related to what i said to the voice bot. if my sentence were too long or complex the voice bot didnt understand so it was necessary to be quite stilted in conversation"

"He asked me to repeat myself 3 times."

Specificity of responses. Based on the replies that VA had provided, some participants perceived that the conversation was very specific and realistic. For example, consider the responses of two participants on the specificity of VA's responses:

"Interestingly, I felt like I answered some of the questions without a typical response. The VA's response still made a lot of sense and was warm."

"They were more specific and more realistic."

In contrast, some participants perceived VA to be responding in a script-like manner offering generic responses in a script-like fashion. For example, a few participants shared the following opinions:

"It seems its answers is rather generic to any response."

"The answers are pretty generic."

Effect of COVID-19. While most participants reported no change in their lives over the course of the study due to the pandemic, a few had lost jobs and loved ones. A few expressed feeling isolated and lonely. Some participants have also mentioned feeling more tired than before. A few participants also mentioned a second lockdown being imposed in the UK during the course of the study.

6. Discussion

Rapid advancements in machine-learning have enabled a widespread usage of voice-based conversational user interfaces in personal and business domains. While they make our lives easier by assisting with daily tasks, they can also be developed to help cater to our emotional needs. In distressing times like the COVID-19 pandemic, they can improve mental resilience and alleviate feelings of loneliness. One route to resilience is by stimulating self-compassion.

The increasing ease of development, scalability and accessibility of voice-based conversational user interfaces make them a prime candidate for a computerised therapist. Further, the perceived non-judgemental nature of conversational user interfaces aids in emotional disclosure and can combat the stigma associated with mental health-care.

While most of the existing literature on conversational user interfaces for mental health-care focuses on treating illnesses like depression (Fitzpatrick et al., 2017; Fulmer et al., 2018), few studies have focused on using conversational user interfaces in a preventative approach (Lee et al., 2019; van As, 2019). The aforementioned studies used text as the modality to interact with the conversational user interface.

The current study focused on preventative mental health-care. It investigated the different roles a voice-based conversational user interface can partake in a conversation to improve self-compassion to provide or ask for care and stimulate self-compassion as an effect. Further, the main focus was to investigate the effects of voice-based interactions because speech is a better medium to facilitate emotional disclosure (Esterling et al., 1994).

The participants volunteered to interact with a voice-based conversational user interface called 'VA' that provides or asks for care in the experimental conditions. In the control condition, the participants engaged in self-compassion stimulating tasks by responding to text-based questionnaires. The study was conducted in three stages, with each stage corresponding to a specific self-compassion stimulating task. Each stage of the study was conducted on a different day. Self-Compassion and perceived loneliness were measured at the beginning and the end of the study.

Based on the results, no statistically significant changes in self-compassion and loneliness scores were observed as an effect of time over the total sample. The effect on self-compassion and loneliness scores due to the interaction between time and condition were also analysed to be statistically not significant. The analysis of changes in self-compassion scores calculated per participant also did not yield any statistically significant results. The equivalence tests show that the effect sizes corresponding to self-compassion scores are equivalent to zero for every condition. This means that in every condition, there is no change in self-compassion. Hence, the relative effectiveness of the conditions on self-compassion scores cannot be determined.

Considering the results of this study, one could suggest that the construct of self-compassion is more of a trait characteristic than a state characteristic, making it difficult to be manipulated in such short interactions. However, the findings of Lee et al. (2019) and van As (2019) contradict the above-mentioned statement. It is crucial to note that the interactions in this study were short-lived and distributed across three different days.

The lack of any statistically significant findings could be attributed to four reasons. First, the amount of time the participants engaged in interacting with VA was very short. In the study by Lee et al. (2019), the participants interacted with the chatbot Vincent for two weeks. In the study by van As (2019), the participants engaged in a single interaction with VA for about 20 minutes. Meanwhile, the participants engaged in three short interactions of 5 minutes each in this study. Further, the Dialogflow API audio quota limits restricted the amount of time a participant could talk to VA per dialogue. The quota also limited the duration of interaction for each stage. Emotional disclosure often requires recollection and retrospection of one's experiences. It is likely that the limited duration of interaction hindered the participant from fully engaging with VA.

Second, long periods of no interaction with VA between each stage of the study could have dampened the study's efforts to increase self-compassion. The study was paused for a week after the first stage to address a problem with VA's technical setup. After troubleshooting, only one stage of either care-receiving and care-giving conditions could be conducted at a time. As a result, each participant was subjected to at least three days

of no interaction with VA between stages. To run the conditions in a parallel fashion as much as possible and avoid other confounds, the control condition was conducted alongside the experimental conditions. This led to considerable delays between the subsequent stages in control condition.

Third, the measures concerning participants' perception of VA and the conversation indicated that they did not experience a delightful interaction with VA. Based on the qualitative analysis, we know that some participants experienced problems concerning VA misunderstanding their responses or asking them to repeat their response multiple times. This could have negatively affected the participants' engagement in the self-compassion stimulating task.

Finally, it is likely that loneliness induced by the lockdown in the UK had a much stronger negative effect on the participants' self-compassion. Though depression, anxiety and loneliness are connected, the mechanism that can combat loneliness through self-compassion by addressing the inter-related symptoms needs further research. While the study by Akin (2010) suggested that encouragement of self-compassion can potentially reduce loneliness, they also mentioned that it is difficult to give a full explanation related to causality between self-compassion and loneliness as their study used correlational data.

While the study focused on engaging the participants in a conversation with a voice-based conversational user interface that was expected to be perceived as gender-ambiguous, very few participants have perceived it to be so. VA was perceived to be more female in the care-giving condition and more male in the care-receiving condition. This is in line with gender stereotypes of females being more care-giving and nurturing than males, but now in reversed causality. In other words, any agent acting in a care-giving manner is perceived more likely to be female, and an agent (in need of) receiving care is perceived more likely to be male.

Two stages of the care-receiving and the care-giving conditions were focused on stimulating self-compassion along the dimensions of common humanity and mindfulness. It was expected that self-compassion would increase, at least along these dimensions. Nevertheless, there were no statistically significant effects observed. The lack of any change

in the common humanity and mindfulness scores can be attributed to the previously mentioned reasons that led to no statistically significant changes in the self-compassion and loneliness scores.

To summarise, this study's findings suggest that short interactions followed by a long duration of no interaction with a voice-based conversational user interface lead to no change in self-compassion. Further, VA was perceived as female in the care-giving condition in line with stereotypical gender bias. Studies with more prolonged and frequent interactions with an engaging voice-based conversational user interface need to be conducted to understand the temporal stimulation of self-compassion better. Further, research has to be done to investigate the causal mechanism between stimulating self-compassion and alleviating perceived loneliness.

Limitations and Future work

The study faced many limitations, mainly concerning the consistency and availability of resources and technical setup. First, the participants were not truly randomly sampled. Participant recruitment was conducted in a two-stage fashion. Participants were recruited first for the experimental conditions and then for the control condition. This was done because of limitations on the Prolific platform that enable recruiting participants for a study across different conditions only in a stage-wise process. It should also be noted that the experimental and control conditions differed significantly in their execution. Therefore, two separate recruitment calls had to be placed on Prolific.

Further, these separate recruitment calls enabled the participants to decide if their technical setup or environment allows them to participate. While the execution of different conditions of the study ran parallel to an extent, the participants were not recruited simultaneously. The absence of true random sampling could likely have led to a sampling bias.

Second, the study had a high drop-out rate which led to the asymmetric sample sizes across conditions. This is due to both poor conversation repair and problems with the participant recruitment platform. While poor conversation repair led to a few participants

dropping out of the study, not all participants received the invitation to subsequent stages of the study. The researchers were made aware of this issue only when some participants informed that they had not received an invitation to the next parts of the study. While repeated invitations were sent, not all participants received them as the Prolific recruitment platform experienced some problems due to an unexpected surge in online studies due to COVID-19.

Third, the participants did not get to interact with VA for a prolonged duration. The duration of each dialogue with VA was time-bound to 25 seconds, a limitation of hosting VA as third-party integration. The participants interacted with VA in three short interactions lasting only 5 minutes each. Each stage was limited to only 5 minutes to accommodate the initially estimated sample size while keeping the Dialogflow API's audio processing quota limitations in mind. While this decision seemed like a fair trade-off, it could have weakened the strength of the manipulation. Further, the prolonged periods of no interaction between the participant and VA between the subsequent stages in each condition could have dampened the manipulation.

Fourth, conversational user interfaces, in general, need a large bank of training data to allow them to respond accurately. Such large training data is typically gathered over prolonged interactions with the conversational user interfaces, and the data correspond to specific tasks. Though VA was intended to be perceived as emotionally intelligent, it was not because its training data was somewhat limited due to time availability and lack of training data suitable for emotionally rich open-ended interactions. The lack of suitable training data led to problems on two fronts. One, VA was not equipped to handle unexpected participant responses leading to misunderstandings sometimes. Two, VA's responses seemed less specific and led to lower engagement with the participants.

Fifth, VA was developed based on the ES configuration of Dialogflow agents. This meant that VA could handle only simple task-oriented linear conversation flows, not the intricacies of an open-ended conversation flow. Since conversations with VA were more emotional, having an underlying state machine model to navigate different conversation states would lead to uninterrupted and pleasurable conversation experiences.

The current offering of voice-based conversational user interface design training within the human-computer interaction domain is somewhat limited, and new frameworks and design guidelines need to be developed (Murad & Munteanu, 2020). It should be stressed that even though the recent advancements in natural language processing have led to the widespread usage of conversational user interfaces, they still need further development to follow general high-level conversations. The current study faced quite a few challenges in designing a voice-based conversational user interface to handle emotional conversations, be perceived as emotionally intelligent, and improve self-compassion.

Based on this study's experiences, it is recommended that future studies randomly sample their participants to avoid any sampling bias. Further, research using voice-based conversational user interfaces developed on a state machine model should be encouraged to handle conversation repair better and not confound the experimental manipulation. Future research could also consider investigating the temporal aspects of self-compassion stimulation using voice-based conversational user interfaces. Investigating the frequency and minimum duration of voice-based interactions required to stimulate self-compassion could help determine the strength of manipulation needed. Further, the challenges associated with having an engaging open-ended conversation with a voice-based conversational user interface could help design self-compassion stimulating tasks where the voice user interface can ask for care. This way, the challenges could be used to design self-compassion stimulating tasks instead of confounding the manipulation.

7. Conclusion

The current study investigated the effect on self-compassion and loneliness due to multiple short-term interactions with a gender-ambiguous Voice User Interface called VA that provided or asked for care.

It was expected that the self-compassion in the participants would increase when they interacted with VA. To test this, the participants volunteered to engage in three short conversations with VA that provided or asked for care in the experimental conditions. Each interaction with VA was conducted on a different day and lasted about 5 minutes. In

the control condition, the participants answered three self-help questionnaires with tasks to stimulate self-compassion. Each questionnaire was made available to the participants on a different day.

The first main takeaway was that no condition led to any significant improvement in self-compassion nor reduction in loneliness. Short interactions followed by long breaks between subsequent interactions could have dampened the strength of experimental manipulation. Further, poor conversation repair likely hurt participants' perception of VA and affected their engagement in the self-compassion stimulating tasks.

The second takeaway was that VA was perceived to be more male in the care-receiving condition and more female in the care-giving condition. This is in accordance with the gender stereotypes of females being more care-giving and nurturing than males, but now in reversed causality - VA acting in a care-giving manner is more likely to be perceived as female, and VA (in need of) receiving care is more likely to be perceived as male.

Addressing the limitations of the current study, we suggest that future studies provide ample time to facilitate emotional disclosure and also opt for a state machine model for the conversation flow. Future studies could also consider the challenges of having open-ended high-level conversations with voice user interfaces as potential building blocks for creating self-compassion stimulating tasks where the voice user interface can ask for care.

References

Akin, A. (2010). Self-compassion and loneliness. *International Online Journal of Educational Sciences*, *2*(3).

Angular team. (2016). *Angular.* Retrieved from `https://angular.io/docs`

Armin Ronacher. (2010). *Flask.* Retrieved from `https://flask.palletsprojects.com/en/1.1.x/`

Asher, S. R., & Paquette, J. A. (2003). Loneliness and peer relations in childhood. *Current directions in psychological science*, *12*(3), 75–78.

Barbato, G., Barini, E., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*, *38*(10), 2133–2149.

Bathke, A. C., Schabenberger, O., Tobias, R. D., & Madden, L. V. (2009). Greenhouse–geisser adjustment and the anova-type statistic: cousins or twins? *The American Statistician*, *63*(3), 239–246.

Benoit Chesneau. (2010). *Gunicorn.* Retrieved from `https://gunicorn.org/`

Berguno, G., Leroux, P., McAinsh, K., & Shaikh, S. (2004). Children's experience of loneliness at school and its relation to bullying and the quality of teacher interventions. *The qualitative report*, *9*(3), 483.

Bérubé, C., Schachner, T., Keller, R., Fleisch, E., von Wangenheim, F., Barata, F., & Kowatsch, T. (2020). Voice-based conversational agents for the prevention and management of chronic and mental conditions: A systematic literature review. *JMIR Preprints 30/11/2020*, *25933*.

Bondevik, M., & Skogstad, A. (1998). The oldest old, adl, social network, and loneliness. *Western Journal of Nursing Research*, *20*(3), 325–343.

Bors, M. L. (2018, Mar). *What is a finite state machine.* Retrieved from `https://medium.com/@mlbors/what-is-a-finite-state-machine-6d8dec727e2c`

Braun, V., & Clarke, V. (2012). Thematic analysis.

Comiteau, L. (2021, feb). *Coronavirus is hitting freedom-loving dutch teenagers hard.* Retrieved from

https://www.dutchnews.nl/features/2021/02/coronavirus-is-hitting-freedom-loving-dutch-teenagers-hard/

Coomes, K. (2018, jun). *From j.a.r.v.i.s to john legend, here are our favorite a.i. assistants.* Retrieved from https://www.digitaltrends.com/home/the-best-ai-assistants/

Corrigan, P. (2004). How stigma interferes with mental health care. *American psychologist*, *59*(7), 614. doi: 10.1037/0003-066X.59.7.614

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811–817.

Donovan, E., Rodgers, R. F., Cousineau, T. M., McGowan, K. M., Luk, S., Yates, K., & Franko, D. L. (2016). Brief report: Feasibility of a mindfulness and self-compassion based mobile intervention for adolescents. *Journal of adolescence*, *53*, 217–221. doi: 10.1016/j.adolescence. 2016.09.009

Esterling, B. A., Antoni, M. H., Fletcher, M. A., Margulies, S., & Schneiderman, N. (1994). Emotional disclosure through writing or speaking modulates latent epstein-barr virus antibody titers. *Journal of consulting and clinical psychology*, *62*(1), 130. doi: 10.1037/0022-006X.62.1.130

Falconer, C. J., Rovira, A., King, J. A., Gilbert, P., Antley, A., Fearon, P., ... Brewin, C. R. (2016). Embodying self-compassion within virtual reality and its effects on patients with depression. *BJPsych Open*, *2*(1), 74–80. doi: 10.1192/bjpo.bp.115.002147

Falconer, C. J., Slater, M., Rovira, A., King, J. A., Gilbert, P., Antley, A., & Brewin, C. R. (2014, 11). Embodying compassion: A virtual reality paradigm for overcoming excessive self-criticism. *PLOS ONE*, *9*(11), 1-7. Retrieved from https://doi.org/10.1371/journal.pone.0111933 doi: 10.1371/journal.pone.0111933

Finlay-Jones, A., Kane, R., & Rees, C. (2017). Self-compassion online: A pilot study of an internet-based self-compassion cultivation program for psychology trainees. *Journal of Clinical Psychology*, *73*(7), 797–816. doi: 10.1002/jclp.22375

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017, Jun 06). 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), e19. Retrieved from `http://mental.jmir.org/2017/2/e19/` doi: 10.2196/mental.7785

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of hci. *interactions*, *24*(4), 38–42.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018, Dec 13). Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Ment Health*, *5*(4), e64. Retrieved from `http://mental.jmir.org/2018/4/e64/` doi: 10.2196/mental.9782

Google. (2016). *Google dialogflow.* Retrieved from `https://cloud.google.com/dialogflow`

Hays, R. D., & DiMatteo, M. R. (1987). A short-form measure of loneliness. *Journal of personality assessment*, *51*(1), 69–81.

Kirby, J. N. (2017). Compassion interventions: The programmes, the evidence, and implications for research and practice. *Psychology and Psychotherapy: Theory, Research and Practice*, *90*(3), 432–455. doi: 10.1111/papt.12104

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355–362.

Leary, M. R., Tate, E. B., Adams, C. E., Batts Allen, A., & Hancock, J. (2007). Self-compassion and reactions to unpleasant self-relevant events: the implications of treating oneself kindly. *Journal of personality and social psychology*, *92*(5), 887.

Lee, M., Ackermans, S., van As, N., Chang, H., Lucas, E., & IJsselsteijn, W. (2019). Caring for vincent: A chatbot for self-compassion. In *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–13). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3290605.3300932` doi: 10.1145/3290605.3300932

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's

only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94 - 100. Retrieved from http://www.sciencedirect.com/science/article/pii/S0747563214002647 doi: https://doi.org/10.1016/j.chb.2014.04.043

MacBeth, A., & Gumley, A. (2012). Exploring compassion: A meta-analysis of the association between self-compassion and psychopathology. *Clinical psychology review*, *32*(6), 545–552.

Miguel Grinberg. (2018). *Flask - socket.io.* Retrieved from https://github.com/miguelgrinberg/Flask-SocketIO

Mullins, L. C., & Dugan, E. (1990). The influence of depression, and family and friendship relations, on residents' loneliness in congregate housing. *The Gerontologist*, *30*(3), 377–384.

Munn, H. (2020, apr). *Why the global recovery from coronavirus depends on mental health research.* Retrieved from https://www.mqmentalhealth.org/posts/why-the-global-recovery-from-coronavirus-depends-on-mental-health-research

Murad, C., & Munteanu, C. (2020). Designing voice interfaces: Back to the (curriculum) basics. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–12).

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 72–78).

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and Identity*, *2*(3), 223-250. Retrieved from https://doi.org/10.1080/15298860309027 doi: 10.1080/15298860309027

Neff, K. D., & Germer, C. K. (2013). A pilot study and randomized controlled trial of the mindful self-compassion program. *Journal of clinical psychology*, *69*(1), 28–44. doi: 10. 1002/jclp.21923

Neff, K. D., & Germer, C. K. (2018). *The mindful self-compassion workbook: A proven way to accept yourself, build inner strength, and thrive.* New York, NY: Guilford.

Neff, K. D., Kirkpatrick, K. L., & Rude, S. S. (2007). Self-compassion and adaptive psychological functioning. *Journal of Research in Personality*, *41*(1), 139 - 154. Retrieved from http://www.sciencedirect.com/science/article/pii/S0092656606000353 doi: https://doi.org/10.1016/j.jrp.2006.03.004

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparld: an r software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical software*, *50*(12).

Perlman, D., & Peplau, L. A. (1982). *Loneliness: A sourcebook of current theory, research, and therapy.* New York, NY: Wiley.

Pinquart, M., & Sorensen, S. (2001). Influences on loneliness in older adults: A meta-analysis. *Basic and applied social psychology*, *23*(4), 245–266.

Pounder, J., & Barton, R. M. . S. (2016). *Humanity in the machine.* Retrieved from https://www.mindshareworld. com/sites/default/files/MINDSHARE HUDDLE HUMANITY MACHINE 2016 0.pdf

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from https://www.R-project.org/

Richardson, T., Elliott, P., & Roberts, R. (2017). Relationship between loneliness and mental health in students. *Journal of Public Mental Health.*

Ryan, M. C., et al. (1998). Hospitalized elderly. *Journal of gerontological nursing*, *24*(3), 19–27.

Sample, I. (2020, dec). *Covid poses 'greatest threat to mental health since second world war'.* Retrieved from https://www.theguardian.com/society/2020/dec/27/covid-poses-greatest-threat- to-mental-health-since-second-world-war

Seligman, M. E., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology progress: empirical validation of interventions. *American psychologist*, *60*(5), 410.

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? comparison of eliza with modern dialogue systems. *Computers in Human Behavior*, *58*, 278–

295.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, *26*(5), 559–569.

Spurgeon, J. A., & Wright, J. H. (2010). Computer-assisted cognitive-behavioral therapy. *Current psychiatry reports*, *12*(6), 547–552. doi: 10.1007/s11920-010-0152-4

van As, N. (2019). *(master thesis)a brief encounter with vincent: the effect on self-compassion from a single interaction with a chatbot that gives or asks for help.*

Weeks, D. J. (1994). A review of loneliness concepts, with particular reference to old age. *International Journal of Geriatric Psychiatry*.

Weiss, B., Wechsung, I., Kühnel, C., & Möller, S. (2015). Evaluating embodied conversational agents in multimodal interfaces. *Computational Cognitive Science*, *1*(1), 6. doi: 10.1186/s40469-015-0006-9

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

West, D. A., Kellner, R., & Moore-West, M. (1986). The effects of loneliness: a review of the literature. *Comprehensive psychiatry*, *27*(4), 351–363.

Xiang, Y.-T., Yang, Y., Li, W., Zhang, L., Zhang, Q., Cheung, T., & Ng, C. H. (2020, apr). Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. *Lancet Psychiatry*, *7*(4), 228–229. Retrieved from `https://doi.org/10.1016/s2215-0366(20)30046-8` doi: 10.1016/S2215-0366(20)30046-8

Yarnell, L. M., Stafford, R. E., Neff, K. D., Reilly, E. D., Knox, M. C., & Mullarkey, M. (2015). Meta-analysis of gender differences in self-compassion. *Self and Identity*, *14*(5), 499–520.

Appendix

1. Sample size estimation

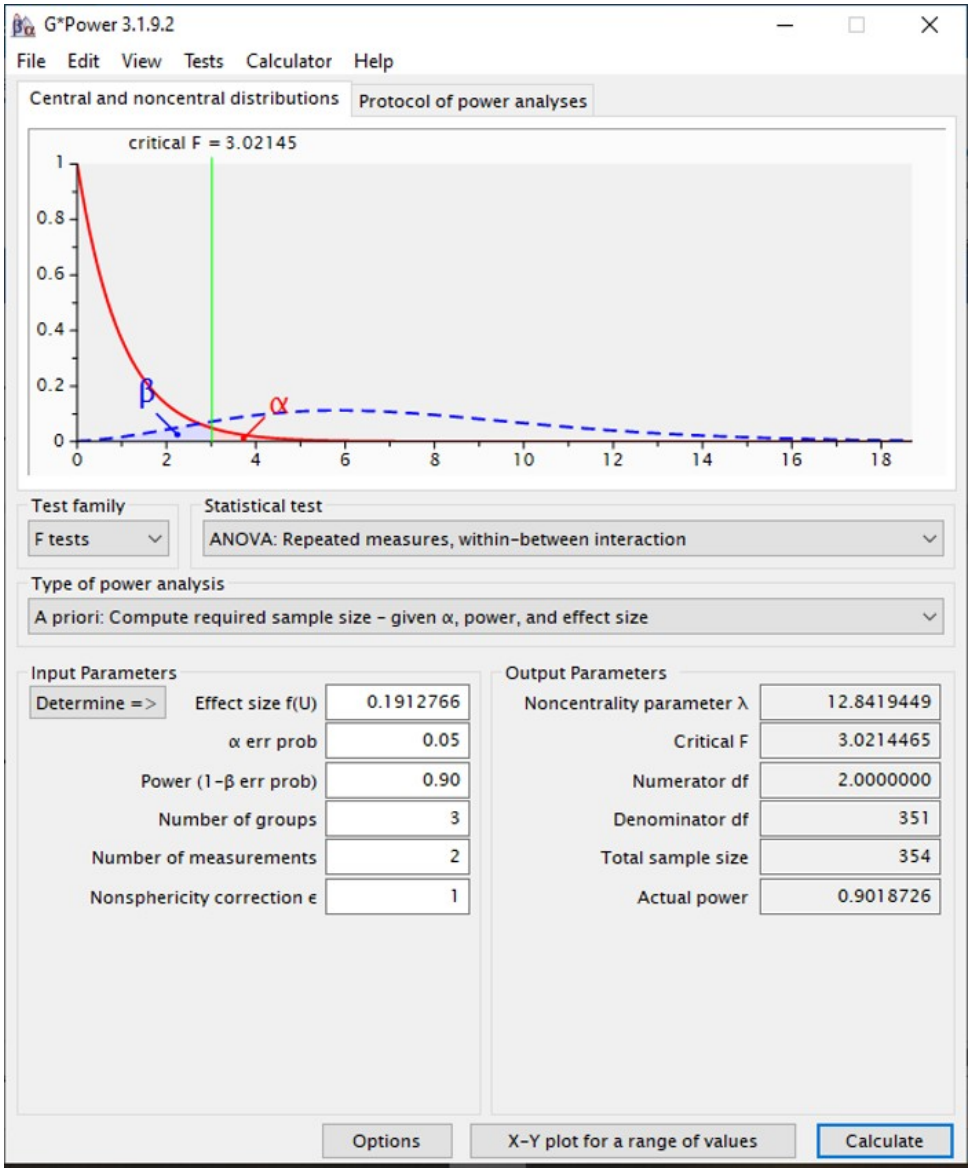

Figure 7. Sample size estimation using G*Power