Design and Evaluation of Virtual Human Mediated Tasks for Assessment of Depression and Anxiety

Joy O. Egede[*]
joy.egede@nottingham.ac.uk

Shashank Jaiswal[*]
shashank.jaiswal@nottingham.ac.uk

Maria J. Galvez Trigo[*]
maria.galveztrigo@nottingham.ac.uk

Dominic Price[*]
dominic.price@nottingham.ac.uk

Natasha Elliot[†]
natasha.elliott@nottingham.ac.uk

Neil Nixon[†]
neil.nixon@nottingham.ac.uk

Deepa B. Krishnan[†]
deepa.krishnan@nottshc.nhs.uk

Richard Morriss[†]
richard.morriss@nottingham.ac.uk

Peter Liddle[†]
peter.liddle@nottingham.ac.uk

Christopher Greenhalgh[*]
chris.greenhalgh@nottingham.ac.uk

Michel Valstar[*]
michel.valstar@nottingham.ac.uk

ABSTRACT

Virtual human technologies are now being widely explored as therapy tools for mental health disorders including depression and anxiety. These technologies leverage the ability of the virtual agents to engage in naturalistic social interactions with a user to elicit behavioural expressions which are indicative of depression and anxiety. Research efforts have focused on optimising the human-like expressive capabilities of the virtual human, but less attention has been given to investigating the effect of virtual human mediation on the expressivity of the user. In addition, it is still not clear what an optimal task is or what task characteristics are likely to sustain long term user engagement. To this end, this paper describes the design and evaluation of virtual human-mediated tasks in a user study of 56 participants. Half the participants complete tasks guided by a virtual human, while the other half are guided by text on screen. Self-reported PHQ9 scores, biosignals and participants' ratings of tasks are collected. Findings show that virtual-human mediation influences behavioural expressiveness and this observation differs for different depression severity levels. It further shows that virtual human mediation improves users' disposition towards tasks.

CCS CONCEPTS

• **Human-centered computing → User studies**; • **Applied computing → Health informatics**; • **Computing methodologies → *Artificial intelligence*.**

KEYWORDS

virtual humans, embodied conversational agents, ECAs, depression, anxiety, mental health

ACM Reference Format:
Joy O. Egede, Shashank Jaiswal, Maria J. Galvez Trigo, Dominic Price, Natasha Elliot, Neil Nixon, Deepa B. Krishnan, Richard Morriss, Peter Liddle, Christopher Greenhalgh, and Michel Valstar. 2021. Design and Evaluation of Virtual Human Mediated Tasks for Assessment of Depression and Anxiety . In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478361

1 INTRODUCTION

Mental health problems are one of the most prominent global challenges facing healthcare systems. This has further been exacerbated by the socioeconomic deprivations imposed by the current COVID-19 crisis such as movement restrictions, loss of income, loneliness, uncertainties, and bereavements leading to increased incidences of mental health disorders such as depression and anxiety. Tackling mental health problems is indeed key to a global economic health recovery, and The World Health Organization's (WHO) Mental Health Action Plan 2013-2020 [36] advocates that improved health information systems would play a vital role in achieving this goal.

Major depression, which very often co-occurs with anxiety, is considered one of the leading causes of disability in developed nations [10]. Its symptoms include low mood, feeling of guilt, loss of interest, poor concentration and low energy. Current treatment methods usually involve prolonged counselling and therapy sessions with a clinician. However, these have been affected by the reduction in face-to-to face meetings due to the COVID pandemic, and are also expensive, repetitive, laborious, highly subjective and time-consuming. It is therefore necessary to develop new methods which are objective and require minimal human intervention to support clinical provisions.

The quest for efficient objective methods have fostered research into virtual technologies that exploit changes in established biomarkers such as stress levels [26], head movements [1, 18], psychomotor symptoms [29], facial expressions [33] , voice inflexion [30] and personality traits [6, 7] for early detection and treatment of mental

[*] School of Computer Science, University of Nottingham, UK
[†] School of Medicine, University of Nottingham, UK.

IVA '21, September 14–17, 2021, Virtual Event, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8619-7/21/09…$15.00
https://doi.org/10.1145/3472306.3478361

health disorders. Recent advances in social signal processing and affective computing have enabled automatic measurements of many aspects of these biomarkers of human behaviour, and these have been utilised for objective prediction of depression and anxiety levels, with little human intervention [32]. Further, [31] demonstrated that the performance of such predictive systems on mental health assessment depends on what task a user is performing whilst being observed. However, there is still an open question of what an optimal task is and what characterises an optimal task.

This work hypothesises that a key factor to obtaining good behavioural signals for depression and anxiety assessment is to use social settings and that Embodied Conversational Agents (ECAs), also referred to as virtual humans (VHs), could provide this social component so that users behave just like they would in normal interaction. To this end, this work presents the design and evaluation of virtual human-mediated tasks for depression and anxiety assessment in a user study involving 56 participants. Half of the participants completed a set of tasks guided by a virtual human, while the other half used a system with no virtual human. Audio-visual and physiological signals were recorded of participants while completing the tasks, together with self-reported depression and anxiety levels using the Patient Health Questionnaire (PHQ9) and Generalised Anxiety Disorder Assessment (GAD-7) respectively. Broadly, the study aimed to investigate the impact of virtual human-mediation on behaviour expressivity during task completions and on users' evaluation of the tasks.

The contribution of this work is fourfold: (i) it show that VH mediation produces stronger visual cues compared to when no VH is used, thus supporting the value of VH in simulating naturalistic social interactions suitable for assessing mental health from behaviour signals; (ii) it demonstrates the capability of the proposed computer-based tasks in eliciting distinctive levels of behaviour expressiveness across depression severity classes, which would be valuable in building real-time predictive systems; (iii) it empirically shows that users' conception of a task's 'attractiveness' at face value, differs significantly after actual engagement with the task, emphasising the importance of active user involvement in digital health technology design and (iv) it highlights the effect of virtual human-mediation on users' evaluation of the computer-based tasks.

2 RELATED WORK

The health sector has witnessed a significant increase in the use of virtual human technology interventions for managing and treating medical conditions. Virtual humans have been used in a broad range of medical applications, including acting as virtual patients in a teaching and learning environment for medical students to practice on [23], playing the role of care assistants to aged populations [16] and delivering health information [25]. Virtual humans are typically equipped with human characteristics such as voice recognition, natural language processing, empathy, emotion recognition and dialogue management to enable their use as user interfaces in applications requiring human-like interaction. Their capabilities vary from simple scripted content delivery [25] enabled by Speech to Text/Text-to-Speech technologies, to more complex Artificial Intelligence (AI) driven naturalistic interactions [2].

As the incidence of mental health disorders has surged in recent times, researchers are focusing more attention on the use of virtual human technologies to facilitate early detection and management of depression and anxiety [12, 15, 20, 21]. These virtual agents take advantage of the social signal processing technologies capable of extracting behaviour primitives (such as eye gaze, voice properties, facial actions and head movements) that encode characteristics associated with mental health disorders. In [11], *SenseiKiosk*, a fully automated virtual human interviewer, is used for assessment of audio-visual indicators of PSTD, depression and anxiety. The ECA engages the user in a dialogue similar to a therapist while exhibiting appropriate behavioural responses. Similarly, [21] explores the concept of *virtual agent as a service* with *Conversagent*, a Low-Intensity Cognitive Behavioural Therapy (LiCBT) coach that supports self-management of moderate levels of depression. Taking this further, [15] investigated the reliability of using a virtual human in mental health assessment. Comparison of the VH administration of the PHQ9 questionnaire to both a self-administration and clinician's administration of the same to a user, showed negligible differences in reported scores.

Virtual human (VH) technologies offer significant advantages in digital healthcare applications. The natural language communication capabilities of ECAs make them suitable interfaces for populations who may otherwise be confused by complicated conventional software interfaces [16]. Studies [20] have also found that people tend to open up more to virtual humans about mental health issues compared to using anonymised questionnaires or human-human conversation.

Despite the huge potential of VHs for mental healthcare, research in this field is still limited. Existing works have focused on either optimising the social abilities (e.g., speech recognition and dialogue management) and emotional expressiveness (e.g., empathy) of the virtual human to induce trust and natural behaviour of users [5, 11, 12] or evaluating its reliability as an assessment tool [15, 21], whereas less attention has been given to investigating the impact of virtual human mediation on the user's behaviour expressiveness. Furthermore, although the activity performed by a user influences the predictive ability of the automatic depression and anxiety estimation systems [31], it is still not clear what is an optimal task, what task characteristics improve expressivity in VH simulated settings, or what tasks users are more likely to engage with. Consequently, this work seeks to examine the following questions: (i) how does VH-mediation influence behaviour expressivity; (ii) how does this expressivity differ across depression severity levels? (iii) What kind of tasks are users most disposed to, and does VH-mediation influence this disposition?

3 VIRTUAL HUMAN MEDIATED TASK (VHT) SYSTEM DESIGN

The VHT System consisted of two main parts: the ECA manager and the Sensory data module, as shown in Figure 1.

ECA Manager: This was developed in Unity 3D [1] as a multi-platform desktop system for running the User study. The ECA Manager is based on the ARIA VALUSPA platform [2], which uses the

[1] https://unity.com/

[2] https://aria-agent.eu

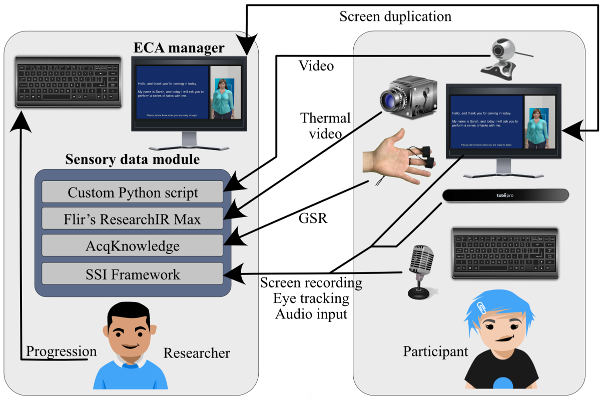
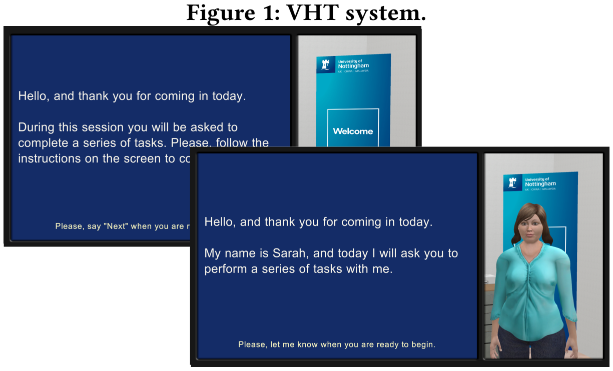


Figure 1: VHT system.



Hello, and thank you for coming in today.

During this session you will be asked to complete a series of tasks. Please, follow the instructions on the screen to co

Please, say "Next" when you are r

Welcome

Hello, and thank you for coming in today.

My name is Sarah, and today I will ask you to perform a series of tasks with me.

Please, let me know when you are ready to begin.

Figure 2: ECA Manager's interface. Text-only guided mode to the left. Virtual-human guided mode overlapping to the right.

Greta ECA [3] and can deliver pre-scripted text, computer-generated speech through the Text-To-Speech engine CereVoice [4], and a visual embodiment of a human with synchronised facial and bodily gestures. To achieve more fluid body animations for the ECA, a series of Mixamo [5] animations were modified in Unity 3D and assigned to the ECA, instead of the animations generated by Greta.

The system's User Interface comprises a window divided in two sections: left and right. In the left section, the tasks that participants are asked to complete during the study are shown, alternating with the text introducing each task and any instructions needed to complete them. In the right section, a banner is shown with or without an ECA in front of it. For the completion of the tasks, the system has two modalities: virtual-human based and text-only based (See Figure 2).

In the virtual-human guided mode, an ECA guides the participant through the different tasks, giving verbal instructions. Whilst the participant is completing a task for which interaction with the ECA is not necessary, the ECA is still present on screen performing a series of gestures or movements (e.g., nodding, breathing) depending on the task on hand. In the text-only guided mode, there is no ECA, instead, all instructions and guidance for completing the tasks are given in written form, with no verbal instructions either. To progress through the different tasks, the participant is prompted to give a verbal command specified by the ECA, if present, as well as written on screen. In order to have a higher level of control on the interaction flow, the researcher running the session would then trigger the progression via the keyboard.

Sensory Data Module: IIn order to understand how the completion of each task and the interaction with the ECA affects each participant, a collection of data-capture devices was used to capture data during participant sessions. The data-capture devices and corresponding software applications used were:

- Video from a Teledyne Dalsa Genie Nano C2050 with Tamron M118FM08 lens with data captured by a custom Python recording script.
- Audio from a Yeti microphone captured with the SSI [35] framework.
- Screen recording captured with the SSI framework.
- Thermal (infra-red) video from a Flir A655sc thermal camera captured with Flir's ResearchIR Max software.
- Galvanic skin resistance (GSR) from a BIOPAC MP160 device (electrodes connected to the participant's index finger) recorded to AcqKnowledge.
- Eye-tracking from a Tobii Pro Fusion captured with the SSI framework.

Recordings from all devices except audio, screen and eye-tracking recordings were synchronised by starting recording simultaneously with an external hardware trigger. The timestamp that the trigger was activated was recorded in SSI so that SSI recordings could be synchronised with the other sources post-experiment.

4 DESCRIPTION OF TASK TYPES

Depression and anxiety have been found to alter a person's normal audio-visual and physiological responses, e.g., increased restlessness for anxiety and decreased facial and vocal activity for depression [24, 37]. These behavioural changes have been exploited for automatic recognition of mental health states via various types of human-computer interaction tasks [34]. This study aimed to design virtual-human guided computer-based tasks (CBTs) that were quick, appealing and engaging to encourage use without a physician's prompting, executable on a mobile device and capable of eliciting behaviour characteristics relevant for digital mental health assessment. These CBTs were designed in collaboration with mental health experts to ensure their appropriateness for the task. Four categories of computer-based interactive tasks were developed for the VHT System, namely: *Mimicking*, *Dyadic Interaction*, *Digital treatment* and *Psychometric*, each consisting of two sub-tasks. These tasks and their relation to clinical mental health management are described below.

4.1 Mimicking Tasks

These tasks aim to induce facial and vocal activity and are popularly used in automatic prediction of depression severity [34].

4.1.1 Sustained Vowel Phonation: asks the user to pronounce the letter A for a duration of at least 6 seconds under four conditions: i) Regular speech volume ii) Loud pronunciation iii) Soft pronunciation iv) Pronunciation while smiling.

[3] https://github.com/isir/greta
[4] https://www.cereproc.com/en/products/sdk
[5] https://www.mixamo.com/

4.1.2 *Facial Expression Mimicking:* shows the user a set of short video clips and then asks the user to mimic/imitate the facial expressions in the clip as closely as possible.

4.2 Dyadic Interaction Tasks

These tasks aim to mirror the patient-clinician interactions in psychotherapy, a mental health treatment approach [9], while maintaining natural expressiveness.

4.2.1 *Emotion Recall:* asks the user to recall two past emotional events, one positive and one negative, followed by questions on how the participant has been feeling during the current COVID-19 crisis and the resultant lockdown and restricted social interaction. The user will be answering these questions to the virtual human agent, which will be programmed to respond appropriately to the user, for example, with empathy (e.g., *"...that must have been a difficult time for you..."*) when the participant describes an event related to a negative emotion [14]. Each emotional event narration takes 2mins, while the questions relating to COVID-19 take 1min, with 1min resting pause between questions.

4.2.2 *Thematic Apperception Dixit Cards:* shows the user a set of images selected from Dixit [6] cards. These images are somewhat *"dream-like"* or abstract in nature and are open to individual interpretation. After each image is shown, the participant is asked to describe their interpretation of the story depicted in the image.

4.3 Digital Treatment Tasks

4.3.1 *Mindfulness:* aims to make the user practice mindfulness techniques by watching a video on the screen. This is hinged on recent reports of the efficacy of mindfulness for depression treatment when integrated with Cognitive Behaviour Therapy [13].

4.3.2 *Reading text aloud:* asks users to read a passage taken from "Harry Potter" so as to assess vocal activity levels.

4.4 Psychometric Tasks

4.4.1 *Emotional Stroop:* originally proposed by [4], this task asks the user to name the ink colour of words presented to them. There are two categories of words presented to the participant: neutral words (e.g., bottle, sky, watch, etc.) and emotional words (e.g., kill, cancer, war, etc.). The response time in naming the colours is logged by the VHT system for behaviour analysis. It is hypothesised that depressed people will take longer to name the colours of words related to negative emotion/depression compared to neutral words.

4.4.2 *Emotional Faces Go No-go task:* is popularly used to measure behavioural inhibition. Here, users are presented with a series of *"go"* and *"no-go"* cues. The aim of the task is to respond as quickly as possible when a *"go"* cue is presented and inhibit response when a *"no-go"* cue is presented. The emotional go no-go task is a modified version of the original task such that the letters or pictorial stimuli commonly used in the original setup are replaced by affective stimuli (e.g., happy faces [go] vs sad faces [no go]). This task has been widely used to test emotional processing in both healthy people and people with affective disorders [28].

5 VHT STUDY PROTOCOL

Prior to the study, participants were asked to fill out five online forms consisting of the *Patient Health Questionnaire (PHQ9)* and *Generalised Anxiety Disorder Assessment (GAD-7)* which both measure severity of anxiety and depression; the *Big Five Inventory (BFI)* for measuring personality traits and style; the *Negative Attitude towards Robots Scale (NARS)* [22] which measures emotions and attitudes that could prevent people from interacting with conversational robots (in this case, the VH); and a COVID-19 risk assessment form, which checks that participants did not have COVID-19 symptoms. The *PHQ9* and *GAD-7* responses will serve as groundtruth reference data for automatic anxiety and depression assessment from behaviour cues.

To be eligible for the study, participants had to be 18+years, fluent in English, not currently undergoing treatment for depression or anxiety, and not have a PHQ9 total score $\geqslant 20$, or a score of 2 on the last question (suicide item) of the PHQ9 form, due to lack of professional mental health support at the study site and ethical considerations. Submitted forms were screened by the researchers, and ineligible persons were informed of this and directed to where to get help.

The study took place in a research lab within the University. Informed consent was first collected from the participant before engaging in any activity. The study session involved participants completing four interactive computer-based tasks, one from each category listed in Section 4, using either of two system modes: the *virtual-human guided mode* or the *text-only guided mode* described in Section 3. This is to allow comparison of participant's experience and behaviour in a simulated social setting (using the VH), to when no VH is used. An alternating scheme was used to assign system modes to participants.

Each participant took part in two study sessions each spaced by one week. In the first session, participants were allowed to choose two tasks while the other two were allocated by the researcher. This was because the study aimed to determine what task characteristics participants found attractive. It was hypothesised that these *favoured* tasks were more likely to sustain long-term engagement, especially as previous studies [21] have reported a decline in users' engagement over time with similar VH mediated systems. In the second session, two tasks were repeated from the first session, with two new ones added. The researcher allocated all tasks in the second session, balancing for task uptake. Participants used the same system mode for both sessions. Each session lasted about 45mins. While completing the tasks, participant's video, audio, thermal (infrared) video, skin conductance, eye-gaze and heart rate data were recorded.

At the end of the study, participants completed an online feedback form which included the following:

- *User Experience Questionnaire (UEQ):* which assesses experience during Human-Computer Interaction under the dimension of attractiveness, novelty, efficiency, dependability, stimulation and perspicuity [17].
- *Technology Acceptance Model (TAM) Questionnaire:* which measures user's tendency to adopt a new technology under the themes of perceived value, perceived ease of use, attitude towards using and intention to use [19]. This was completed

[6] https://www.libellud.com/dixit/

only by participants who used the virtual-human guided mode.

- *Task Assessment Questionnaire:* rates each completed task on a 5-point Likert scale, from 1(least preferred) to 5(most preferred). It also included open-ended questions to understand the reason for the assigned ratings and collect general feedback on the system.

This work focuses on the analysis of the *PHQ9*, *GAD*, Task Assessment Questionnaire and video data.

6 DATA ANALYSIS AND FINDINGS

6.1 Participants

A total of 56 participants took part in the study: 33 females and 23 males, all within the ages of 18 to 45. Participants' self-reported PHQ9 scores P_s ranged from *minimal* to *moderately severe* depression. Note that two persons (not part of the 56), with $P_s > 20$, i.e., *severe* depression, were excluded from the study following the eligibility criteria but were given information on how to get help. Half of the participants (i.e., 28) used the virtual-human guided mode, while the other half used the text-only guided mode.

Figure 3 shows the PHQ9 score and age distribution of participants. The *minimal* (19) and *mild* (22) groups made up a greater proportion of the data, followed by *moderate* (11), while *moderately severe* (4) had the least representation. Only participants who were not currently undergoing treatment for depression were allowed to take part, hence the high proportion of the lower PHQ9 groups. Also, most of the participants were students recruited from within the University, since the COVID-19 movement restrictions limited participation from outside the University. This led to the significantly larger number of 18-25yrs age group in the dataset.

6.2 Influence of the System mode on Users' assessment of Tasks

Each participant used only one system mode (i.e., virtual-human guided or text-only guided mode) and rated the tasks on completion of the study. Thus, it was necessary to check if the system mode influenced users' assessment of the tasks. For each system mode, the mean user ratings per task were extracted from the data. Further, T-tests were applied for each task to establish if the mean differences observed between the two modes were merely due to chance. Figure 4 shows the comparison of the mean ratings for each task categorised by system mode. Significant differences (at $p = .05$) were found between the system modes in task ratings for *Facial Expression Mimicking (.04)* , *Emotion Recall (.01)*, *Thematic Apperception (.03)* and *Mindfulness (.04)* tasks, with mean ratings for the virtual-human guided mode being significantly higher than the text-only guided mode, except for the *ER* task, where ratings were lower for the VH-guided mode.

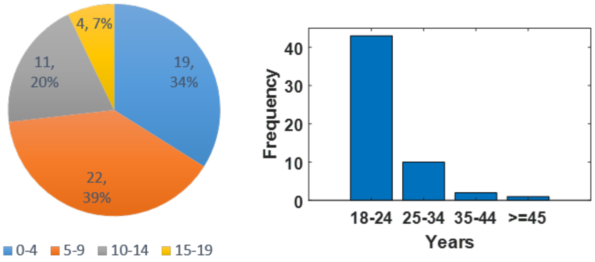


Figure 3: Distribution of Participants' self-reported PHQ9 Scores (left) and Age (right) in the dataset.

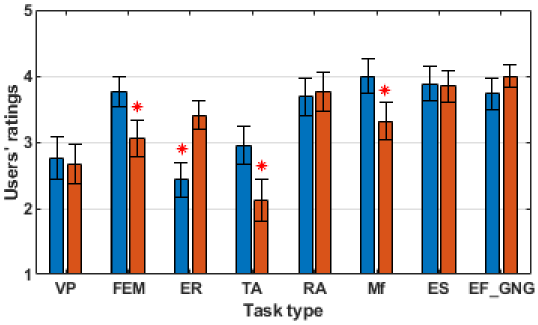


Figure 4: Comparison of Users' Task ratings categorised by the type of system used. Significant differences are indicated with * at *p=0.05*. Tasks are defined as FEM: Facial Expression Mimicking; VP: Vowel Phonation; ER: Emotion Recall; TA: Thematic Apperception; RA: Reading aloud; Mf: Mindfulness; ES: Emotional Stroop; EF_GNG: Emotional Face Go-No-go.

The results imply that the virtual human interactivity improved the users' disposition towards the *FEM*, *TA* and *Mf* tasks whereas, users preferred the *ER task* in the text-guided mode. The outcome on the *ER task* was quite unexpected as it was hypothesized that *ER* being a conversational task would benefit most from VH-mediation. This was also contrary to findings in [20] where virtual humans were found to increase self-disclosure of mental health issues, compared to filling out a written questionnaire; although in their study, the VH's rendition of the questionnaire was modified to promote interaction, which could have helped disclosure. In contrast, in this study, the conversations were kept identical for both the VH & text-guided modes. Further analysis of the qualitative data provided by participants regarding their task preferences would be required to understand this observation.

6.3 Users' Evaluation of the Computer-based Tasks

A system-mode independent comparison of user ratings for each task was conducted to determine which tasks were most preferred. The mean ratings for each task were computed across all participants. Figure 5 presents a comparison of the participants' preference ratings for each task. Looking at average ratings by task categories, it can be observed that the psychometric (3.92) and digital treatment (3.71) tasks were preferred by most users compared to the mimicking (2.93) and dyadic (2.89) tasks. Assessed individually, the Emotional Face Go-No-go task is most preferred, while the Thematic Apperception task is least preferred.

Further analysis was done to find out which tasks participants thought attractive prior to experiencing it and if this influenced their subsequent ratings of the tasks. Figure 6 shows the frequency of selection of tasks in the first study session.

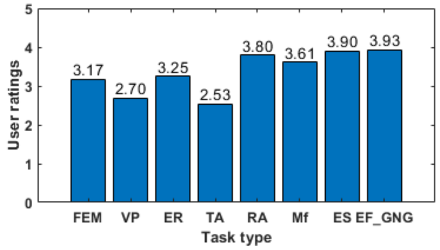


Figure 5: Comparison of task ratings independent of the System mode used. Task acronyms are defined as in Figure 4.

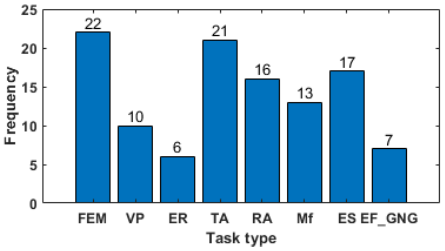


Figure 6: Comparison of Task selection frequency in participants' first sessions. Task acronyms are defined as in Fig. 4.

The results show that, within the four task categories of mimicking, dyadic interaction, digital treatment and psychometric, the *FEM*, *TA*, *RA* and *ES* task respectively, were more appealing to participants at face value. Conversely, participants were less likely to choose Emotion recall and the Emotional Faces Go-No-go (EF_GNG) tasks. Some possible explanations for the aversion to ER task are the avoidance of the additional effort of thinking up a positive and negative emotional experience to discuss; sharing such in the presence of others (researchers) [11], or not having a fitting experience to share. The low take-up for the Go-No-go task could have been due to the restriction of selecting only one task from each category, which would then imply that the Stroop task held more attraction to participants.

Looking at the task selection data in the context of the task ratings in Figure 5, the *EF-GNG* task achieves a higher rating than the FEM and TA tasks which seem to have been most selected by participants. In fact, the TA task has the lowest ratings (2.53) across tasks compared to Go-No-Go (3.93), although its high uptake could have been due to avoidance of the ER task. This suggests that task appeal at face value does not necessarily translate to user preference of the task.

6.4 Influence of VH-mediation on visual behavioural expressivity

To determine whether the use of the virtual human had an impact on participants' behavioural expressiveness, a comparative analysis of users of the two system modes was conducted on two behavioural components: *facial activity* and *degree of head movements*. These behavioural cues were extracted from videos recorded of participants while interacting with the systems.

6.4.1 Head movement descriptors. These were defined as changes in head pitch, roll and yaw and were extracted using OpenFace 2.0 [3]. OpenFace outputs head pose measurements per frame. Since the length of the videos varied and data was generated per frame, video descriptors were generated using statistical measures and custom video level metrics. The statistical measures of *mean*, *variance* and *range* were explored as they better encode the magnitude of displacement compared to other statistical measures.

6.4.2 Facial activity descriptors. These were defined as the activations of upper and lower face muscles (also referred to as facial action units (FAUs)). Eighteen FAU occurrences and 17 FAU intensities were extracted for each video frame. Action unit intensities range from 0 (not active) to 5 (Highest activation), while the occurrences are scored as 0 (absent) or 1 (present). For the action unit intensities, the same statistics defined for the head movement data were used as video descriptors. In addition, since neutral expressions constitute a greater proportion of the video frames, a custom video descriptor P_t was created and defined as the proportion of active video frames with intensities above the mid-level activation (2.5); the aim was to determine which group exhibited stronger facial activity. For the facial action unit occurrences which were categorical data, a custom video descriptor P_{active}, expressed as percentage values, was defined as the proportion of video frames for which a facial action unit is active.

6.4.3 Behaviour Comparison of User groups. T-tests were conducted to compare the behavioural expressivity observed between the VH and non-VH system users. Each video descriptor was explored as the dependent variable in the T-tests, while the user grouping was the independent variable in all cases. Each user group consisted of 28 participants. The results of the tests are shown in Table 1. Note that for brevity, only behavioural activity and descriptors for which statistically significant differences were found between the groups are reported here.

On head movement data, the T-test comparison showed differences between the two groups on both head yaw and roll but not for pitch, as shown in Table 1. Comparing the means for each group in terms of the *Range* statistics, the results suggest that the VH guided system users made wider/broader head movements than those who used the text-only guided system. Similarly, in terms of the variance, the results suggest more movements away from the central head position during the sessions involving the VH.

For the facial data, in terms of intensity of expressions, stronger activations of lower face muscles -–*AU15 (lip corner depressor)* and *AU20 (lip stretcher)* — were observed for the VH mode compared to the no-VH mode. In terms of frequency of facial activity, more upper face activity –AU7 (lid tightener) – was observed in the VH-mode, whereas more frequent lower facial activity–*AU17 (chin raiser)* and *AU20 (lip stretcher)* –was observed for the no-VH mode. However, although *AU20* occurred more often in the no-VH mode, its activations were stronger in the VH mode.

These findings could mean that VH system users found the interaction more human-like, and as such were more expressive as opposed to reading instructions off the screen. Regarding the relatively higher head yaw (left-right head turn) observed in the VH group, it is also possible that system design artefacts could have influenced head activity, e.g., looking from the on-screen text to the virtual human, but this needs to be confirmed empirically with further analysis of the eye gaze data.

Table 1: Behaviour comparison of VH versus non-VH System Users. AU definitions in order of appearance: AU26, AU15, AU20, AU7 and AU17. Head pose data is measured in radians; P_{active} and P_t are percentage values.

Behaviour	Descriptor	VH System $N = 28$	NVH System $N = 28$	*p-value* ($\alpha = .05$)
FAU intensity				
AU15	P_t	**0.10**	0.03	.04
AU20	P_t	**0.04**	0.01	.04
FAU Occurrence				
AU7	P_{active}	**26.93**	18.13	.004
AU17	P_{active}	8.65	**11.37**	.004
AU20	P_{active}	24.51	**26.9**	.036
Head Pose				
Yaw	*Range*	**1.671**	1.385	.005
Roll	*Range*	**1.624**	1.259	.0032
Yaw	*Variance*	0.01	0.007	.017

Table 2: Comparison of behaviour expressiveness between the minimal, mild and moderately severe PHQ9 groups. Head pose data is measured in radians. The super-scripted letters indicate PHQ9 groups pairs with statistically significant differences.

Behaviour	Descriptor	Minimal $N = 22$	Mild $N = 19$	Moderate $N = 15$	*p-value* ($\alpha = .05$)
Facial activity	P_{active} *(All)*	22.39^a	22.43^b	25.5^{ab}	$.019^a$, $.021^b$
Yaw	*Mean*	0.082^a	0.089	0.102^a	.044
Roll	*Mean*	0.067^a	0.072	0.087^a	.024

The performance of detection systems for depression and anxiety levels has advanced lately due to their improved ability to detect distinctive verbal & non-verbal behaviour associated with these conditions [31]. VH-mediated systems capable of evoking these characteristics could help fast-track the development of personalised digital interventions that would promote treatment accessibility to a broader population. They could potentially provide some level of privacy for people who refrain from seeking medical help due to perceived negative socio-cultural and economic implications [8]. Further, with patients' consent, it could provide therapists with information of the patient's progress in-between psychotherapy sessions or to help assess the efficacy of pharmacological treatments.

6.5 Influence of PHQ9 scores on visual behaviour expressivity

Preliminary analysis was performed to investigate whether there was an effect or not of self-reported PHQ9 scores on levels of visual behaviour display. The dataset consisted of four categories of self-reported PHQ9 scores (See Section 6.1). To compare the expressiveness of the PHQ9 groups, pairwise T-tests were conducted. The categories of *moderately severe* and *severe* were combined into one group for the analysis, due to the small sample size of the latter.

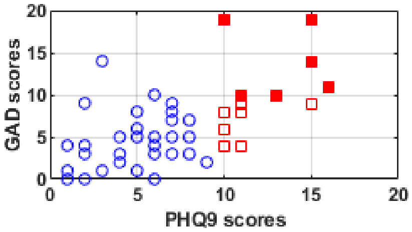


Figure 7: Participants' GAD7 vs PHQ9 scores. *PHQ9* ≥ *10* & *GAD7* ≥ *10* indicate *moderate* depression and *moderate* to *severe* anxiety respectively.

For the head pose data analysis, the same video level descriptors described in section 6.4 were used. For the facial activity, a single metric combining all action unit activations was computed for each participant. Specifically, the average of P_{active} descriptor was taken across FAUs and denoted as $P_{active}(All)$. Table 2 presents the results of the pairwise T-tests for the PHQ9 groups.

The results show that the *moderately* severe PHQ9 groups exhibited significantly more facial activity compared to the mild and *minimal* groups, but there was no significant difference between the *minimal* and *mild* groups. Similarly, the *moderate* group showed more head activity than the *minimal* group but no significant difference to the *mild*. No difference was found between the *mild* and *minimal* groups either.

The higher incidence of head activity in the moderate group appears to reflect anxiety behaviour (i.e. restlessness) [37] rather than depression, as severely depressed people display little or no facial and head movement [24]. This observation was further investigated in Fig 7, which shows a plot of the PHQ9 scores of participants against their GAD7 scores. It can be seen that 40% of the *moderate* depression group indicated by the solid red squares also reported *moderate* to *severe* anxiety scores, which could have been responsible for the activity pattern observed in this group. Both conditions often coexist, and in some cases, depression is considered an offshoot of progressed anxiety [27]. However, these findings are limited by the combination of the *moderate* and *moderately severe* group, due to the small sample size. More insight would be gained by a four-class comparison with sufficient representations.

7 CONCLUSION AND FUTURE WORK

This work has described the development of a virtual human-mediated task system for mental health assessment and its evaluation in a User study. Participants' audio-visual data and self-reported PHQ9 scores were recorded. Analysis of the video data revealed that the VH guided mode elicited stronger behavioural displays from users compared to the text-only guided mode, even though the latter produced more frequent but less intense facial activity, thus establishing the value of virtual humans in supporting mental health management. It further showed that the moderately depressed groups exhibited more visible activity compared to less depressed groups, an outcome attributed to anxiety. Lastly, it showed that the VH significantly affected users' attitude towards the tasks. Future work will focus on video analysis of behaviour per task, to determine which ones are more informative for mental health assessment, and also look at the qualitative data to gain more insight into users' perception of both the tasks and the mode of completion.

ACKNOWLEDGMENTS

The research reported in this paper was conducted by the NIHR Nottingham Biomedical Research Centre. This work has been funded by the National Institute for Health Research. The views represented are the views of the authors alone and do not necessarily represent the views of the Department of Health in England, NHS, or the National Institute for Health Research.

REFERENCES

[1] Andreas Altorfer, Stefan Jossen, Othmar Würmle, Marie-Louise Käsermann, Klaus Foppa, and Heinrich Zimmermann. 2000. Measurement and meaning of head movements in everyday face-to-face communicative interaction. *Behavior Research Methods, Instruments, & Computers* 32, 1 (2000), 17–32.

[2] ARIAVALUSPA. 2020. *ARIA VALUSPA*. Retrieved May 7, 2021 from https://aria-agent.eu/

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, New York, NY, USA, 59–66.

[4] Moshe Shay Ben-Haim, Paul Williams, Zachary Howard, Yaniv Mama, Ami Eidels, and Daniel Algom. 2016. The emotional Stroop task: assessing cognitive performance under exposure to emotional content. *Journal of visualized experiments: JoVE* 112 (2016).

[5] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling* 59, 1 (2005), 21–30.

[6] O Joseph Bienvenu and Murray B Stein. 2003. Personality and anxiety disorders: a review. *Journal of Personality disorders* 17, 2: Special issue (2003), 139–151.

[7] Mina Brandes and O Joseph Bienvenu. 2006. Personality and anxiety disorders. *Current psychiatry reports* 8, 4 (2006), 263–269.

[8] Evelien PM Brouwers. 2020. Social stigma is an underestimated contributing factor to unemployment in people with mental illness or mental health issues: position paper and future directions. *BMC psychology* 8 (2020), 1–7.

[9] Joseph K Carpenter, Leigh A Andrews, Sara M Witcraft, Mark B Powers, Jasper AJ Smits, and Stefan G Hofmann. 2018. Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. *Depression and anxiety* 35, 6 (2018), 502–514.

[10] European Commission. 2005. Improving the mental health of the population: Towards a strategy on mental health for the European Union.

[11] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.

[12] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e19.

[13] Simon B Goldberg, Raymond P Tucker, Preston A Greene, Richard J Davidson, David J Kearney, and Tracy L Simpson. 2019. Mindfulness-based cognitive therapy for the treatment of current depressive symptoms: a meta-analysis. *Cognitive behaviour therapy* 48, 6 (2019), 445–462.

[14] Russell Harris. 2006. Embracing your demons: An overview of acceptance and commitment therapy. *Psychotherapy in Australia* 12, 4 (2006).

[15] Shashank Jaiswal, Michel Valstar, Keerthy Kusumam, and Chris Greenhalgh. 2019. Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 81–87.

[16] Patrick Kenny, Thomas Parsons, Jonathan Gratch, and Albert Rizzo. 2008. Virtual humans for assisted health care. In *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*. 1–4.

[17] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, 63–76.

[18] Stuart John Leask, Bert Park, Priya Khana, and Ben DiMambro. 2013. Head movements during conversational speech in patients with schizophrenia. *Therapeutic advances in psychopharmacology* 3, 1 (2013), 29–31.

[19] Younghwa Lee, Kenneth A Kozar, and Kai RT Larsen. 2003. The technology acceptance model: Past, present, and future. *Communications of the Association for information systems* 12, 1 (2003), 50.

[20] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 4 (2017), 51.

[21] Martin H Luerssen and Tim Hawke. 2018. Virtual Agents as a Service: Applications in Healthcare. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 107–112.

[22] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies* 7, 3 (2006), 437–454.

[23] Andrew B Raij, Kyle Johnsen, Robert F Dickerson, Benjamin C Lok, Marc S Cohen, Margaret Duerson, Rebecca Rainer Pauly, Amy O Stevens, Peggy Wagner, and D Scott Lind. 2007. Comparing interpersonal interactions with a virtual human to those with a real human. *IEEE transactions on visualization and computer graphics* 13, 3 (2007), 443–457.

[24] Robert E Rakel. 1981. Differential diagnosis of anxiety.

[25] Albert A Rizzo, Belinda Lange, John G Buckwalter, Eric Forbell, Julia Kim, Kenji Sagae, Josh Williams, Barbara O Rothbaum, JoAnn Difede, Greg Reger, et al. 2011. *An intelligent virtual human system for providing healthcare information and support*. Technical Report. MADIGAN ARMY MEDICAL CENTER TACOMA WA.

[26] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676.

[27] Robert A Schoevers, DJH Deeg, W Van Tilburg, and ATF Beekman. 2005. Depression and generalized anxiety disorder: co-occurrence and longitudinal patterns in elderly patients. *The American Journal of Geriatric Psychiatry* 13, 1 (2005), 31–39.

[28] Kurt P Schulz, Jin Fan, Olga Magidina, David J Marks, Bella Hahn, and Jeffrey M Halperin. 2007. Does the emotional go/no-go task really measure behavioral inhibition? Convergence with measures on a non-emotional analog. *Archives of Clinical Neuropsychology* 22, 2 (2007), 151–160.

[29] Christina Sobin and Harold A Sackeim. 1997. Psychomotor symptoms of depression. *American Journal of Psychiatry* 154, 1 (1997), 4–17.

[30] Cynthia Solomon, Michel F Valstar, Richard K Morriss, and John Crowe. 2015. Objective methods for reliable detection of concealed depression. *Frontiers in ICT* 2 (2015), 5.

[31] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. 2020. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing* (2020).

[32] Siyang Song, Linlin Shen, and Michel Valstar. 2018. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 158–165.

[33] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 3–10.

[34] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 3–10.

[35] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. 2013. The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*. 831–834.

[36] WHO. 2020. *Mental disorders*. Retrieved May 6, 2021 from https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[37] Andrea Stevenson Won, Brian Perone, Michelle Friend, and Jeremy N Bailenson. 2016. Identifying anxiety through tracked head movements in a virtual classroom. *Cyberpsychology, Behavior, and Social Networking* 19, 6 (2016), 380–387.