Original research

DIGITAL HEALTH

Digital Health
Volume 8: 1–12
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076221139092
journals.sagepub.com/home/dhj
SAGE

Assessing medical students' perceived stress levels by comparing a chatbot-based approach to the Perceived Stress Questionnaire (PSQ20) in a mixed-methods study

Julia-Astrid Moldt[1], Teresa Festl-Wietek[1], Amir Madany Mamlouk[2] and Anne Herrmann-Werner[1,3]

Abstract

Objective: Digital transformation in higher education has presented medical students with new challenges, which has increased the difficulty of organising their own studies. The main objective of this study is to evaluate the effectiveness of a chatbot in assessing the stress levels of medical students in everyday conversations and to identify the main condition for accepting a chatbot as a conversational partner based on validated stress instruments, such as the Perceived Stress Questionnaire (PSQ20).

Methods: In this mixed-methods research design, medical-student stress level was assessed using a quantitative (digital- and paper-based versions of PSQ20) and qualitative (chatbot conversation) study design. PSQ20 items were also shortened to investigate whether medical students' stress levels can be measured in everyday conversations. Therefore, items were integrated into the chat between medical students and a chatbot named Melinda.

Results: PSQ20 revealed increased stress levels in 43.4% of medical students who participated ($N = 136$). The integrated PSQ20 items in the conversations with Melinda obtained similar subjective stress degree results in the statistical analysis of both PSQ20 versions. Qualitative analysis revealed that certain functional and technical requirements have a significant impact on the expected use and success of the chatbot.

Conclusion: The results suggest that chatbots are promising as personal digital assistants for medical students; they can detect students' stress factors during the conversation. Increasing the chatbot's technical and social capabilities could have a positive impact on user acceptance.

Keywords

Medical students, stress, communication, mixed-methods design, conversational agent, PSQ20

Submission date: 6 May 2022; Acceptance date: 28 October 2022

Introduction

Background

Increasing digitisation and the shift to digital teaching have created major challenges for universities, teachers and students.[1] As digitisation progresses, the learning environment for university students will also change. In particular, the everyday life of medical students will be influenced not only by an increased range of digital learning content but

---

[1]TIME – Tübingen Institute for Medical Education, University Hospital Tuebingen, Tuebingen, Germany
[2]Institute for Neuro- and Bioinformatics, University of Luebeck, Luebeck, Germany
[3]Department of Internal Medicine VI/Psychosomatic Medicine and Psychotherapy, University Hospital Tuebingen, Tuebingen, Germany

Corresponding author:
Julia-Astrid Moldt, TIME – Tübingen Institute for Medical Education, Elfriede-Aulhorn-Straße 10, 72076 Tuebingen, Germany.
Email: julia-astrid.moldt@med.uni-tuebingen.de

Creative Commons NonCommercial-NoDerivs CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits non-commercial use, reproduction and distribution of the work as published without adaptation or alteration, without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

also by the cooperation of the agents involved. For example, electronic communication via WhatsApp groups and electronic submission systems are already ubiquitous.[2,3] However, the number of personal meetings and face-to-face conversations continues to decrease. Digital learning in isolation poses a risk factor, especially for those who are already depressed or highly stressed.[4] Studies show that medical students in particular have underlying and highly developed stress levels as well as high prevalence rates for mental illness.[5–7] This is also evident in the increasing number of students with stress, anxiety and burnout symptoms.[8–10] Studying medicine places high demands on medical students: large amounts of knowledge must be mastered in a short amount of time.[11] Expectations of high academic performance can increasingly lead to stress during medical education. Increasing stress, in turn, can lead to a decline in academic performance and motivation.[12,13] Previous studies have found a correlation between chronic stress among medical students and increased cynicism and reduced empathy during the course of their studies, which could impact their relationships with patients in the future.[14] In addition, due to distress during their medical education, a lack of integrity and substance abuse have been identified.[13] Considering the high-stress levels and the associated negative impact on both the health of medical students and the quality of their future medical work, the promotion of stress-reducing measures during studies stands to reason.

To address the challenges of education digitisation and the digital learning environment, digital-assistance systems such as chatbots could potentially provide medical students with support in organising their studies. Chatbots could be available to students independent of time and place to support them with questions relevant to their studies. At the same time, chatbots could also act as a personal assistant to whom they could confide about stressful situations. In addition, AI-based approaches were shown to be able to predict the manifestation of major depression based on the conversational behaviour of users, and the use of chatbots also showed a positive effect in already diagnosed depressed patients.[15,16]

In this project, we dedicate ourselves to a chatbot named Melinda (which stands for my emotion-screening learning-supportive integrative new digital assistant).[17] The Melinda project addresses the challenges in changing communication between medical students and lecturers in the context of digital university teaching. It attempts to address an important gap since comprehensive support is increasingly difficult to implement in study-centred asynchronous teaching formats. Previous results have indicated the need for a personal digital assistant embedded with routine chat that can screen for the user's psychosocial needs and react accordingly.[17]

Chatbots as conversational agents for students

The term chatbot is not a clearly defined term in academia; other terms include bots, conversational agents, conversational interfaces and chat agents.[18] What they all have in common is that they are classified as digital dialogue systems, that is software components that simulate human natural language in text form. They start a conversation with users and consumers using text-based language to facilitate human–machine communication. Chatbots can be implemented in two main categories: natural language processing (NLP) chatbots and rule-based chatbots. NLP chatbot systems enable free text input and are therefore usually based on NLP.[19] In essence, NLP chatbots are artificial intelligence (AI) and can act as an intelligent conversational partner; they access text-processing systems and generate answers in text form based on available data.[20] The chatbot Melinda is a rule-based chatbot that is provided with a set of responses guided by a decision tree. Rule-based chatbots use predefined conversational paths and structures and can thus only communicate using a limited knowledge base; they are not based on AI.[21]

Mental-health chatbots and virtual agents for emotional support are already being used in the field of mental health[22] and in learning environments using electronic devices and media.[23] Chatbots can help address organisational tasks, such as providing information, helping find an appointment, counteracting long waiting times or preventing perceived stigma and mental-health barriers in a functional and empathetic way.[24] Examples of such chatbots are Replika,[25] a system that can optionally impersonate a friend, mentor or even a romantic partner, as well as Wysa[24] and Woebot,[15] both of which are digital companions that provide 24/7 exercises and are available for advice or just to talk. Both approaches have been shown in clinical studies to be effective in reducing symptoms of stress or depression.[26,27]

With the Melinda project, we aimed to design a chatbot that can act as a personal digital assistant to medical students, communicate with them in an interested and empathetic way and also measure their stress levels. Accordingly, this article addresses the following two questions:

1. How effective is a conversational bot such as Melinda in assessing the stress levels of medical students and what are the conditions for being accepted as a conversational partner?
2. Can an established measurement tool, such as the Perceived Stress Questionnaire (PSQ20), be transferred into written dialogue to measure student stress levels using the chatbot Melinda? The purpose is to find out if there is a significant difference between the responses of the digital and paper PSQ20 versions.

Methods

Design

We adopted a mixed-methods approach in this study to facilitate the exploration of data and research questions from different perspectives.[28] First, we designed the quantitative part based on PSQ20, a standardised questionnaire measuring actual subjective stress.[29] Second, qualitative analysis was used as a supplementary survey method to analyse the data.[30] We aimed to investigate the extent to which Melinda succeeds in conducting everyday conversations with medical students while assessing their stress levels. In general, chatbot evaluation is challenging due to their complex nature.[31,32] Therefore, we analysed segments of conversations with the chatbot via the International Organization for Standardization (ISO) 9241-11 usability characteristics (i.e., effectiveness, efficiency and satisfaction), which specify the user ergonomics of human–system interaction.[33]

Procedure

To address the research questions, this study was integrated into the regular communication course for medical students. The course was conducted online due to the COVID-19 pandemic. First, students received theoretical input on psychological stress and burnout in medical school and advice on how to cope with mindfulness practices. Then, they were asked to complete exercises on coping strategies for psychological stress, after which they were invited to participate in the study. As a first step, students were asked to complete the validated paper-based PSQ20 version to determine their general stress level. They were then given an abbreviated version of the PSQ20 with selected items integrated into the chat with Melinda. The completion order of the different questionnaire versions was based on our research interest, that is we wanted to measure the stress level of the students before reducing the PSQ20 questionnaire to nine items.

Participants

Participants were medical students in their second year of study. They were invited to take part in the study during their regular communication course in the 20/21 winter term. Participation was voluntary and pseudonymised with a unique code number. The participating students were asked to complete two versions of the questionnaire (paper versus digital). The total number of participants was 136 for the paper version and 148 for the digital version.

Measurements: The perceived stress questionnaire

The subjectively perceived stress load of the medical students was assessed using PSQ20.[29,34] The normal version consists of 30 items. The short version consists of 20 items that can be separated into four subscales (see Table 1). The questionnaire is reliable with a Cronbach's alpha of 0.80–0.86 for the overall score. Questions about perceived stress were evaluated using a four-point Likert scale (almost never, sometimes, often and usually), and participants were asked to rate their subjective perception and evaluation of stressors that affected them in the previous four weeks.[35] By completing the PSQ20 questionnaire of the medical students, a score was calculated for each of the scales and for the overall perceived stress according to the scale calculation of Fliege et al.[36] The sum value can be between zero and 100, with high values representing an increased subjective stress load. Values above the cut-off scores indicate clinically relevant stress.[29,34,36] Table 1 presents the cut-off scores, (sub-)scales and their dimensions for the PSQ20.[37]

Table 1. Subscales, dimensions and cut-off-scores of the PSQ20.

(Sub-)scale	Dimensions	Cut-off score
Worries	Worries, fears about the future and feelings of frustration:	>11
Tension	Exhaustion, imbalance and a lack of physical relaxation	>13
Lack of joy	-	<17
Demands	A Lack of time, deadline pressure or task load	>13
Overall score	Clinically relevant stress	>50

Description: Higher scores indicate higher stress levels. For the subscale *lack of joy*, lower values indicate higher stress levels. Please see the cut-off scores for the definition of higher stress level.[37]

Selection of PSQ20 questions for the chatbot

We used a paper- and digital-based version of the PSQ20 for the chatbot.[36] The Melinda chatbot was first developed at the University of Lübeck, Germany, for a qualification thesis and was further explored for our research study.[38]

Originally, 10 of the 20 PSQ20 questions were intended to be asked in the chatbot conversation to identify whether the answers allow us to draw conclusions about perceived stress prior to the study. Due to a bug in the chatbot's program, one PSQ20 question was (incorrectly) no longer recognised as a PSQ20 question, which is why we could only use nine items. Thus, the subscale *demands* were incomplete with one item missing, and, based on this, we were not allowed to use the subscale *demands* and decided to only report the subscale *lack of joy* to compare the two questionnaire versions. The selected items were integrated into the conversation process for Melinda. For more information about the detailed explanation of the selection, please refer to the qualification paper.[38]

In our study, the subscale *lack of joy* was reliable for both versions of the questionnaire with a Cronbach's alpha of 0.667 for the digital version and 0.640 for the paper version.

Melinda: Bot operation and implementation in a teaching unit for medical students

The conversation content of the decision tree-based bot was based on questions about studying, the COVID-19 pandemic situation and the personal situation and well-being of the students. The chatbot was programmed to provide positive, encouraging, empathic and compassionate responses. For better conversation flow, the nine selected PSQ20 items were rephrased into questions and queried during the chat to fit into the content. The chatbot Melinda is a simple chatbot with a pre-defined conversational flow to facilitate conversation with medical students. The training data consists of 622 individual words (taken from short sentences) that are assigned to one or more intents. After reducing them to their word stems and removing stop-words using the Lancester-Stemmer,[39] a list of 507 tokens is left. Thus, the number of tokens is directly dependent on the size of the training vocabulary, and the vocabulary we used for the initial implementation of Melinda is rather small-sized.[39,40] The intent-classifier we are using here is a multi-layer perceptron (MLP) with ReLUs as an activation function, consisting of a single hidden layer of size 253 and an output layer of 14 classes of intent ('hobby related', 'greeting', 'goodbye', 'age', 'name', 'help request', 'consulting hours', 'sleep-related', 'university', 'internship', 'corona', 'neutral intent', 'negative intent', and 'positive intent'). Our initial conversational manager was designed to talk with medical students about their learning situation during the COVID-19 pandemic. In the context of this study, the 'positive', 'negative', and 'neutral intent' are of particular importance.[41] The network has been trained using a bag-of-words approach[42] and an Adam optimiser[43] with a training length of 5000 epochs, a 0.001 learning rate, betas of 0.9 and 0.999, an epsilon of 10^{-8} and a weight decay of 0.

Table 2 shows nine of 20 PSQ20 items in their original version[36] as well as how we integrated them into the conversation with Melinda. For comparison, the possible answers to the PSQ20 questions were integrated analogously to the original questionnaire with the options 'almost never' through to 'usually'; the rest could be entered as free text. After the students completed the PSQ20, they chatted with Melinda for nine of the 20 items, which were transcribed into written dialogue.

Table 2. Common set of items based on PSQ20.

Data set PSQ20 P PSQ20 items from the paper-based version (original version)	Data set PSQ20 M PSQ20 items transferred and integrated into Melinda chat
(1) You are full of energy.	Do you feel full of energy?
(2) You are light-hearted.	Would you say you are light-hearted?
(3) You feel that too many demands are being made on you.	Do you sometimes have the feeling that the demands are too high?
(4) You feel safe and protected.	Do you feel safe and protected?
(5) You feel under pressure from deadlines.	Do you feel pressured to meet deadlines?
(6) You have too many things to do.	Do you have a lot to do right now?
(7) You feel you are in a hurry.	Do you feel rushed?
(8) You feel you are doing things you really like.	Do you do things you really like?
(9) You enjoy yourself.	Are you having fun?

Study datasets

Statistical analyses were performed by creating two datasets with nine selected items from PSQ20 in its German-modified version. Dataset 1 presents the paper-based version of PSQ20 (PSQ20 P). Dataset 2 provides the chat with Melinda with the selected items integrated digitally into the conversation (PSQ20 M).

Data analysis

All statistical analyses were conducted using IBM SPSS version 27. For the quantitative comparison of the chat and questionnaire answers, the two data sets were coded based on the nine items and checked for usable data.

The elimination criteria included incomplete chat histories that did not contain the PSQ20 or question–answer combinations that did not make sense (for details, please see attachment). After screening according to our exclusion criteria, we obtained complete datasets from the original 136 participants for the paper questionnaire and from the 148 participants of the digital-chat version with Melinda; a total of 41 medical students fully completed both questionnaire versions and were eligible for the final analysis. For the subclasses, we focused on the subcategory *lack of joy*, as it was embedded completely in the conversation trees.

One purpose of this work was to identify whether a significant difference exists in the answers to the digital and paper versions. Indeed, responses could be influenced by adjusting or embedding the questions in the conversation and by embedding. Hence, we calculated the mean, standard deviation and frequencies. The data were normally distributed. The Pearson correlation coefficient was calculated to capture the linear relationships between the relevant variables of each dataset. To examine differences between the means of the two samples, we analysed the statistically significant differences between the distributions of the different PSQ20 versions using the Wilcoxon signed-rank (WSR) test for dependent samples. The significance level in testing this hypothesis was $\alpha = 0.05$. We reject H_0 for $p > \alpha = 0.05$ if the test variable exceeds the critical value z. The critical value z for $\alpha = 0.05$ is $Z\,[1-0.025] = Z\,[0.975] = 1.96$.[44]

The fact that we had to reduce the data for the final analysis to 41 participants who completed both questionnaires suggests that we should also take a closer look at the high error rate of the chatbot. Therefore, we conducted a qualitative exploration of extreme cases in the chat processes in addition to the quantitative analysis of the perceived stress.[45]

Chat analysis was based on thematic analysis, a flexible analysis approach according to Braun and Clarke.[46] For this purpose, a category system was developed based on ISO 9241-11:2018 *Ergonomics of human–system interaction – Part 11: Usability: Definitions and concepts*. Please see the attachment for the used category system.[33] The ISO 0241-11 standard deals with guidelines for human–computer interactions. It contains requirements for the working environment as well as for hardware and software. The aim of these guidelines is to avoid health hazards when working with a product and to make it easier for the user to solve tasks. If the specified measures of effectiveness, efficiency and satisfaction are adequately met, the product has an acceptable level of usability. Effectiveness is typically measured by *task completion* and *accuracy*; the measures of efficiency include *communication effort* and the measures of satisfaction include *ease of use*, *context-dependent questions and learnability* of the chatbot. Standards such as ISO 9241-11 are particularly suited for emerging technologies such as chatbots because they define best practices, are objective and ensure consistency of work.

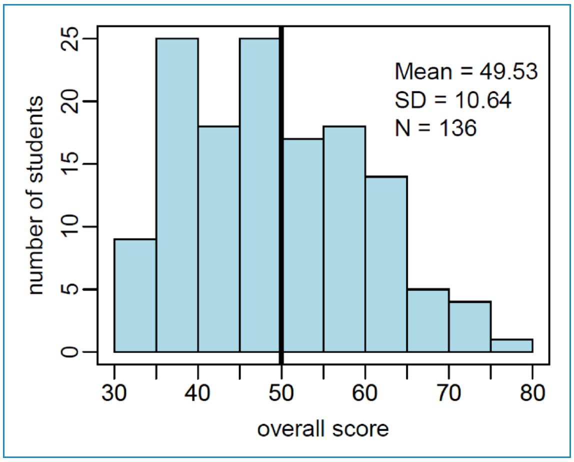

Figure 1. The overall score of the PSQ20 with all items (*x*-axis: PSQ20 overall score; *y*-axis: cumulated number of students, *N* = 136).

The included usability features according to ISO 9241-11 are as follows[33]:

- Effectiveness: The accuracy and completeness with which users achieve a given goal. This means that the tasks at hand are to be performed by a system as completely and correctly as possible.
- Efficiency: The effort used in relation to accuracy and completeness with which users achieve a specific goal. This means that the user must be able to reliably solve the tasks at hand with the existing system functionality with as little effort as possible. Efficiency is related to effectiveness. This can be measured by the time it takes a user to complete a task.
- Satisfaction: This criterion describes the satisfaction of the user when using an interactive system.

Results

PSQ20 for medical students

For students who completed the paper version (N = 136), the mean total score (sum of all 20 items) was 49.53 (SD 10.64), as shown in Figure 1. The range of the overall score was between 30 (lowest) and 76 (highest) points. If the cut-off score is valued (>50), 43.4% of the students have an increased stress load. The evaluation of the subscales also

Table 3. Results of the PSQ20 (sub-)scales above the cut-off score.

(Sub-) scale	Dimensions	Cut-off score	Number of students above the cut-off score (*N* = 136)
Worries	Worries, fears about the future and feelings of frustration:	>11	30.1% (41/136)
Tension	Exhaustion, imbalance and a lack of physical relaxation	>13	92.6% (126/136)
Lack of joy	-	<17	74.3% (101/136)
Demands	A Lack of time, deadline pressure or task load	>13	70.6% (96/136)
Overall score	-	>50	43.4 (59/136)

suggests an increased subjective perception of stress (Table 3). Considering the respective cut-off scores for the subscales, the results in Table 3 were obtained.

Principal results of statistical analysis

For clarity, we limited the descriptive portrayal of the analysis to mean, standard deviation and WSR test results. Table 4 summarises these results. By comparing the means of the nine items of the two datasets, it is evident the students' answers are very similar (Table 4). All items in the two questionnaires underwent Pearson correlation (Table 4) and WSR testing. For this, bivariate pairs were formed for the nine items and for the subscale *lack of joy* for datasets PSQ20 P and PSQ20 M. Pearson's correlation coefficient for each item suggests a significant correlation between the pairs, except for Item 6 (Table 4).

Results of the qualitative study

Students were asked to evaluate the chat with Melinda afterward. Reported positive experiences of chatting with Melinda included that the bot was friendly, the questions were logically structured and the conversation was pleasant. Some communication elements of Melinda worked very well. For example, Melinda welcomed the students and said goodbye, and, in doing so, Melinda achieved a successful personal form that was situationally appropriate and had personalised responses. Melinda also maintained natural appearing conversation using response phrases to develop fluency in conversations.

However, a number of reported negative aspects of the conversation were predominant. For example, the students criticised Melinda's limited ability to answer questions correctly, which is also its largest disadvantage; that is, it has a high error rate. Misinterpretations, answering without a question, duplications and wrong answers made fluent conversation impossible. Additionally, students lacked the necessary transparency as to why they were communicating with the bot and what purpose it served. Some stated they did not see any advantage of the chatbot over a simple questionnaire. Moreover, although some found the bot's conversational style empathetic and natural, many evaluators indicated they found the conversation artificial. The bot's responses seemed exaggerated and overly concerned and compassionate. Some felt the questions were inappropriate because they were questions about personal situations.

We evaluated the thematic analysis of the chats with Melinda according to ISO 9241-11 with the usability characteristics of chatbot effectiveness, efficiency, and satisfaction.[33] See appendix for examples, please note that the conversations with the chatbot were not correct in German either, so we tried to translate this as well as possible to illustrate the problems with the chatbot.

Table 4. Mean, standard deviation and Pearson correlation for nine items between the PSQ20 P and PSQ20 M datasets.

Item	PSQ20 P, $n = 41$, mean (SD)	PSQ20 M $n = 41$, mean (SD)	Pearson correlation	WSR (p-value)
1. You are full of energy.	2.59 (0.81)	2.80 (1.00)	$r = 0.70$; $p < .001$	.06
2. You are light-hearted.	2.54 (0.90)	2.83 (0.95)	$r = 0.58$; $p < .001$	.04
3. You feel that too many demands are being made on you.	2.24 (0.92)	2.00 (0.74)	$r = 0.63$; $p < .001$	.04
4. You feel safe and protected.	3.54 (0.84)	3.51 (0.78)	$r = 0.70$; $p < .05$	.97
5. You feel under pressure from deadlines.	2.56 (1.00)	2.20 (1.00)	$r = 0.75$; $p < .001$	.003
6. You have too many things to do.	2.63 (0.94)	2.85 (0.94)	$r = 0.29$; $p > .05$	.26
7. You feel you are in a hurry.	2.07 (1.06)	2.20 (1.03)	$r = 0.65$; $p < .001$	.40
8. You feel you are doing things you really like.	2.88 (0.84)	2.73 (0.81)	$r = 0.58$; $p < .001$	.22
9. You enjoy yourself.	3.07 (0.72)	3.27 (0.84)	$r = 0.34$; $p < .05$	.17
10. Lack of joy (cut-off) score: < 17.	14.61 (3.00)	15.15 (2.80)	$r = 0.70$; $p < .001$	0.11

The Wilcoxon signed-rank test revealed a significant difference between the two questionnaire versions ($n = 41$) for Item 2 ($Z = -2.03$; $p = .04$), Item 3 ($Z = -2.04$; $p = .04$) and Item 5 ($Z = -2.93$; $p = .003$). Concerning Item 2, students tended to choose lower ratings on the Likert scale for the paper version of the PSQ20 than in the conversation with Melinda. For Items 3 and 5, they tended to answer higher ratings on the Likert scale. No significant differences could be found for any of the other items in the two questionnaires.

- **Effectiveness**: We wanted to know how accurately Melinda conducts conversations with participating students and whether the conversations were completed. We found several parts where the conversation ended unanswered or where Melinda did not understand the student answers and continued following the algorithm. A lack of flexibility was also shown when Melinda did not understand the answers at some parts of the conversation.
- **Efficiency**: We wanted to explore how balanced the conversation between Melinda and the students was regarding proportion and communication effort. When analysing the chat with Melinda, most students answered questions in a very short and brief manner. At some point, Melinda encouraged the students to tell a little bit more, but they did not answer her request, did not reply or negated it. Interestingly and conversely, as students wrote in more detail, Melinda was the one who answered shortly and continued with her algorithm.
- **Satisfaction**: Satisfaction means 'the extent to which the user experience meets the user's needs and expectations'.[33] First, we examined the ease of use based on whether students would chat with Melinda again. Then we examined the extent to which Melinda was able to maintain context-related dialogue. Finally, we examined whether Melinda was able to improve her conversation strategy to answer the students adequately.

The results suggest that a variety of factors are related to a non-satisfying experience with Melinda. The students seemed to frequently lose patience when chatting with Melinda because they responded in a slightly annoying way. This mainly occurred when Melinda's answers did not match the answers of the students or when Melinda repeated a question without understanding the answers. Furthermore, we recognised some bugs in the chat outputs, where the question–answer combination was incorrect. We assumed a transmission error in the written chats. In essence, the chatbot generally succeeded in conducting a conversation, but many (technically related) errors occurred throughout the conversations, which prevented fluent exchange.

Discussion

Principal findings

We used a mixed-methods approach to evaluate the effectiveness of a chatbot in assessing the stress levels of medical students in everyday conversations and to identify the main condition for accepting a chatbot as a conversational partner. Comparing the digital and paper PSQ20 versions, it is evident that the transformation of the PSQ20 items into an everyday conversation with a chatbot was successful. Students showed a general interest in writing with our chatbot, indicating that the use of a chatbot in everyday life could be a suitable (learning) tool for students. Since there were predominantly no significant differences in the responses across both versions, we can assume that the reformulation of the items from a statement to a question (which is more suitable for everyday use) can be well-integrated into a conversation with a chatbot. In the future, a digital study assistant could also obtain information about student stress levels from everyday conversations without explicitly querying the PSQ20 items.

Due to the communication elements that worked well, Melinda was able to recognise conversation beginnings and endings, which are socially and structurally linked pair sequences.[47] Melinda's ability to sustain a conversation by repeating questions help to either positively or negatively reinforce one of the student's statements or engage the student in small talk.[48] Since we wanted to identify the success conditions of future chatbots for students, we focused on the details where the communication fails (see the Appendix for a detailed description of the dialogues). After evaluating the chatbot's response behaviour and the students' feedback concerning their interaction with Melinda according to the categories of effectiveness, efficiency and satisfaction,[33] it is evident that essential elements are still missing for students to accept Melinda as a socially acting chatbot.

Explanations for failed dialogue

There were several well-known reasons why the conversations with Melinda failed. For example, when Melinda lacked the flexibility to respond beyond her set of rules, she revealed herself as not human and highlighted a gap in the expectation of the students that Melinda was not able to fulfil. To understand important aspects of human–machine interaction, it is necessary to have the linguistic behaviour of human–human interaction in mind.[49] Ethnomethodological conversation analysis emerged in the United States in the 1960s; it addresses the analysis of everyday non-purposeful conversations and interactions.[50] The fundamental theory of its founder, U.S. sociologist Harvey Sacks, was that there is 'order at all points' in every communication.[51] Accordingly, every conversation follows a certain pattern of action: the turn-taking system.[52] This means that speech contributions are always contextual; they only refer to the context of their origin or can only be understood in relation to the immediately preceding action.[47] A successful conversation is achieved when the contributions of the participants always mutually and reflexively refer to each other. This implies that intentions, attitudes and interpretations are attuned to the other person and negotiated reciprocally.[47] However, dialogues are sensitive to disruptions because information is constantly renegotiated between the interactors, which is called common ground.[49] If disruptions occur, such as delays, verbal errors or misunderstandings, repairs can be initiated. This can be done by oneself or by the student, for example, by repeating or correcting what has been said.[50,51] Contrary to human beings, who develop their communicative skills over a lifetime in society, AI can only respond to dialogues in chat within the scope of its programming and cannot adequately respond to shared social knowledge.[49] Since most bots cannot access common ground or reliably draw conclusions from what is written, Melinda's awareness of disruptions to the conversation was not possible. Nevertheless, students pursue repair strategies (e.g. repeating sentences) to re-establish common ground, even though they know the system cannot access the same resources as a human counterpart.[53] Besides functionality and performance, social skills, including personality and emotion, are also relevant for interacting with conversation agents.[54,55] Dissatisfaction can be generated in the user if these expectations are not met.[56] As stated by Mori et al., there is an uncanny valley effect regarding human empathy towards robots and other AI entities. Another term for this is the level of affinity, for example, human familiarity increases as chatbots become more human-like in appearance, but, if they become too close to human appearance, there is a risk of eliciting uncanny and uncomfortable feelings and aversive reactions.[57,58] This would explain why some students found it unnatural for Melinda to seem *too* empathetic. Some studies also conclude that bots with more human characteristics, such as those with an animated avatar and voice, are more likely to fall into an uncanny valley compared to bots that are exclusively text-based.[57,59] The bot's social characteristics and features are dependent on the application area and conversational context.[60–62] The most suitable characteristics for the academic field might be a chatbot that encourages students, helps them in stressful situations, provides them with information in a sensitive way,[18] and whose personality is proactive and reliable in guiding and engaging students.[15,17] The foregoing discussion implies that the user, conversational agents and context affect the results of the conversation. The evaluation of the communication behaviour of the chat with Melinda and the assessment of the students' evaluation show various potentials for improvement for the further development of a chatbot. It quickly becomes clear that the exemplary analysis aspects can be largely attributed to limitations in Melinda's programming. Furthermore, there is a need to understand interactional machinery behind human–human service to enable appropriate responses and activity levels to understand the interactional path.[19] This study used ISO-recommended methods to assess effectiveness, efficiency and satisfaction. The methods can help to develop a better understanding of the usability of chatbots and can be adapted for use with medical students.[63]

Limitations

This study was not without some limitations. First, since this is German-language data, the user field will presumably be limited to the German-speaking region. Furthermore, the integration of the PSQ20 items may be inadequately operationalised because factors such as gender, age, response time and conversation duration were not considered. As a result, an in-depth analysis of the participating students was not possible. Limitations regarding the interpretability of the quantitative results are possible since the PSQ20 is based on a total score that makes the subjective perceived stress comparable, which, in turn, is not calculable with our selected nine items. Due to the allocation problems of the bot and the resulting communication difficulties, we had to drastically reduce the amount of analysable data. With further bot development, we will be able to obtain more reliable data.

Conclusion and research perspective

Our results suggest that many students are willing to chat with a bot if it meets the minimum standards to avoid potential frustrations (e.g. ease of use or maintaining conversational flow). However, the chatbot Melinda, which was developed for this study, requires revision to meet these standards. Successful integration of a validated questionnaire, such as the PSQ20, requires careful consideration regarding item wording to ensure that it is consistent with the original content. Assuming it is possible to develop a chatbot that is capable of having a fluid and meaningful conversation, chatbots could be used to assess students' feelings and needs that indicate potential stressors. Students' assessment after conversing with the bot and our qualitative analysis of the chats revealed that willingness and acceptance to talk with a bot are dependent on certain preconditions. In future research, we intend to continue the development of the chatbot by reducing the error rate, which will ensure an effective flow of conversations. Therefore, in the future, the study assistant bot shall be developed in a user-centred way to create a digital infrastructure. The framework conditions of the bot (e.g. possibilities and limitations) shall be known to the students to overcome the expectation gap and to create an environment that deals with personal and health-related topics in a sensitive way. Notwithstanding, it is important to continue to foster research of personal conversation agents in medical education. The data collected with the final version of the bot can be shared with the deep-learning community and for researching study-related stress.

Acknowledgements: We would like to thank our study assistants Marie Thiwissen and Nadja Spannowsky for their help. Further we would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Contributorship: AHW and TFW were responsible for the study design/implementation and data collection. AMM developed the chatbot platform and supported the technical implementation of the chatbot. JAM analysed the research material and conducted the writing for the manuscript. TFW participated in data analysis

and interpretation and critically revised the manuscript. AHW and AMM made significant contributions to the study design and critically revised the manuscript. All authors approved the final version of the manuscript and agreed to be responsible for all aspects found therein.

Ethical approval: The study received ethical approval from the Ethics Committee of Tuebingen Medical Faculty (no. 314/2020BO2). Participation was voluntary, and students provided their written informed consent. All responses and data were kept anonymous.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by the Open Access Publishing Fund of the University of Tuebingen, Deutsche Forschungsgemeinschaft.

Guarantor: JM.

Supplemental material: Supplemental material for this article is available online.

ORCID iD: Julia-Astrid Moldt https://orcid.org/0000-0002-2418-150X

References

1. Aristovnik A, Keržič D, Ravšelj D, et al. Impacts of the COVID-19 pandemic on life of higher education students: a global perspective. *Sustainability* 2020; 12: 8438.
2. Kergel D, Heidkamp B, Telléus PK, et al. *The Digital Turn in Higher Education. Proc International Perspectives on Learning and Teaching in a Changing World*. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-19925-8
3. Iancu AM, Kemp MT and Alam HB. Unmuting medical students' education: utilizing telemedicine during the COVID-19 pandemic and beyond. *J Med Internet Res* 2020; 22: e19667.
4. Breitenbach A. Digital teaching in times of COVID-19: risks and opportunities. *ISA Pedagogy Series* 2021; 1: 1–19.
5. Backović DV, Ilić Živojinović J, Maksimović J, et al. Gender differences in academic stress and burnout among medical students in final years of education. *Psychiatr Danub* 2012; 24: 175–181.
6. Thees S, Gobel J, Jose G, et al. Die Gesundheit von Studierenden im Bologna-prozess. *Prävention und Gesundheitsförderung* 2012; 7: 196–202.
7. Erschens R, Keifenheim KE, Herrmann-Werner A, et al. Professional burnout among medical students: systematic literature review and meta-analysis. *Med Teach* 2019; 41: 172–183.
8. Dyrbye LN, Thomas MR and Shanafelt TD. Systematic review of depression, anxiety, and other indicators of psychological distress among U.S. and Canadian medical students. *Acad Med* 2006; 81: 354–373.
9. Brazeau CM, Shanafelt T, Durning SJ, et al. Distress among matriculating medical students relative to the general population. *Acad Med* 2014; 89: 1520–1525.
10. Santen SA, Holt DB, Kemp JD, et al. Burnout in medical students: examining the prevalence and associated factors. *South Med J* 2010; 103: 758–763.
11. Gusy B, Lohmann K and Drewes J. Burnout bei Studierenden, die einen Bachelor-Abschluss anstreben. *Prävention und Gesundheitsförderung* 2010; 5: 271–275.
12. Park J, Chung S, An H, et al. A structural model of stress, motivation, and academic performance in medical students. *Psychiatry Investig* 2012; 9: 143–149. 20120402.
13. Dyrbye LN, Thomas MR and Shanafelt TD. Medical student distress: causes, consequences, and proposed solutions. *Mayo Clin Proc* 2005; 80: 1613–1622.
14. Thomas MR, Dyrbye LN, Huntington JL, et al. How do distress and well-being relate to medical student empathy? A multicenter study. *J Gen Intern Med* 2007; 22: 177–183.
15. Fitzpatrick KK, Darcy A and Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Health* 2017; 4: e19.20170606.
16. Gamon M, Choudhury M, Counts S, et al. *Predicting Depression via Social Media*. 2013.
17. Herrmann-Werner A, Loda T, Junne F, et al. "Hello, My Name Is Melinda" – students' views on a digital assistant for navigation in digital learning environments; a qualitative interview study. *Frontiers in Education* 2021; 5: 1–7. DOI: 10.3389/feduc.2020.541839.
18. Chaves AP and Gerosa MA. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 2020; 37: 729–758.
19. Bickmore T, Trinh H, Asadi R, et al. Safety first: conversational agents for health care. In: Moore RJ, Szymanski MH, Arar R, et al. (eds) *Studies in conversational UX design*. Cham: Springer International Publishing, 2018, pp.33–57.
20. Hussain S, Ameri Sianaki O and Ababneh N. *A survey on conversational agents/chatbots classification and design techniques*, pp. 946–956. Cham: Springer International Publishing, 2019.
21. Mornhinweg N. Intelligente Chatbots im Gesundheitswesen – Prozessoptimierung im Gesundheitsbereich, von der Dienstleistung bis hin zum Controlling. In: Pfannstiel MA and Straub S (eds) *Dienstleistungscontrolling in Gesundheitseinrichtungen*. Wiesbaden: Springer Fachmedien Wiesbaden, 2018, pp.143–160.
22. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 2018; 25: 1248–1258.
23. Kerry A, Ellis R and Bull S. *Conversational agents in E-learning*, pp. 169–182. London: Springer London, 2009.
24. Inkster B, Sarda S and Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth* 2018; 6: e12106. 20181123.
25. Pardes A. The emotional chatbots are here to probe our feelings, https://www.wired.com/story/replika-open-source (2016, 2016).
26. Sweeney C, Potts C, Ennis E, et al. Can chatbots help support a person's mental health? Perceptions and views from mental

healthcare professionals and experts. *ACM Transactions on Computing for Healthcare* 2021; 2: 1–15.

27. De Nieva JO, Joaquin JA, Tan CB, et al. Investigating Students' Use of a Mental Health Chatbot to Alleviate Academic Stress. *6th International ACM In-Cooperation HCI and UX Conference*. Jakarta &amp; Bandung, Indonesia: Association for Computing Machinery, 2020, pp. 1–10.
28. Mason J. Mixing methods in a qualitatively driven way. *Qual Res* 2016; 6: 9–25.
29. Fliege H, Rose M, Arck P, et al. The Perceived Stress Questionnaire (PSQ) reconsidered: validation and reference values from different clinical and healthy adult samples. *Psychosom Med* 2005; 67: 78–88.
30. Kelle U. Mixed methods. In: Baur N and Blasius J (eds) *Handbuch Methoden der empirischen Sozialforschung*. Wiesbaden: Springer Fachmedien Wiesbaden, 2014, pp.153–166.
31. Casas J, Tricot M-O, Abou Khaled O, et al. Trends & Methods in Chatbot Evaluation. In: *Companion Publication of the 2020 International Conference on Multimodal Interaction* 2020, pp. 280–286.
32. Radziwill NM and Benton MC. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:170404579* 2017.
33. ISO 9241-11. Ergonomics of human-system interaction–part 11: Usability: Definitions and concepts. 2018.
34. Levenstein S, Prantera C, Varvo V, et al. Development of the Perceived Stress Questionnaire: a new tool for psychosomatic research. *J Psychosom Res* 1993; 37: 19–32.
35. Fliege H, Rose M, Arck P, et al. PSQ-Perceived Stress Questionnaire. 2009.
36. Fliege H, Rose M, Arck P, et al. SQ. Perceived Stress Questionnaire [Test description, Scale calculation for PSQ20 and PSQ30 (German), PSQ20 in German, English, German (last 2 years), PSQ30 in German, English, French, Italian, Italian (last month), and Spanish]. Trier: ZPID: Leibniz Institute for Psychology (ZPID) (Ed.) Open Test Archive, 2009.
37. Fliege H, Rose M, Arck P, et al. Validierung des "Perceived Stress Questionnaire" (PSQ) an einer deutschen Stichprobe. *Diagnostica* 2001; 47: 142–152.
38. Buhr M. *Implementierung eines Deep Learning Chatbots zur Analyse des Stresslevels Studierender anhand ihrer Interaktionen mit dem Chatbot*. Luebeck: Universität zu Lübeck, 2021.
39. Paice CD. Another stemmer. *ACM SIGIR Forum* 1990; 24: 56–61.
40. Bird SL, Klein E and Loper E. Natural Language Processing with Python, https://www.nltk.org/ 2009.
41. Agarap AF. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:180308375* 2018.
42. McTear MF, Callejas Z and Griol D. *The conversational interface*. Switzerland: Springer, 2016. https://doi.org/10.1007/978-3-319-32967-3
43. Kingma DP and Ba J. Adam: A Method for Stochastic Optimization. *3rd International Conference for Learning Representations*. San Diego: Cornell University, 2015.
44. Schira J. *Statistische Methoden der VWL und BWL: Theorie und Praxis*. Munich: Pearson Studium Deutschland GmbH, 2009.
45. Denscombe M. Communities of practice. *J Mix Methods Res* 2008; 2: 270–283.
46. Braun V and Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006; 3: 77–101.
47. Forsthoffer I and Dittmar N. Konversationsanalyse. In: Kühl, S., Strodtholz, P., Taffertshofer, A. (eds) Handbuch Methoden der Organisationsforschung. VS Verlag für Sozialwissenschaften. 2009. https://doi.org/10.1007/978-3-531-91570-8_17
48. Picard RW. What does it mean for a computer to "have" emotions. *Emotions in Humans and Artifacts*. M.I.T Media Laboratory Technical Report #534, 2003: 213–235.
49. Lotze N. *Chatbots: eine linguistische Analyse*. Berlin: Peter Lang International Academic Publishers, 2016.
50. Birkner K, Auer P, Bauer A, et al. *Einführung in die Konversationsanalyse*. Walter de Gruyter GmbH: Berlin, 2020.
51. Sacks H. *Lectures on conversation: Volume I*. Malden, Massachusetts: Blackwell, 1992.
52. Sacks H, Schegloff EA and Jefferson G. A Simplest Systematics for the Organization of Turn Taking for Conversation** *Studies in the Organization of Conversational Interaction*. Elsevier, 1978, pp. 7–55.
53. Fischer K. *What Computer Talk Is and Isn't: Human-Computer Conversation as Intercultural Communication*. 2006.
54. Jain M, Kumar P, Kota R, et al. Evaluating and Informing the Design of Chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*. Hong Kong, China: Association for Computing Machinery, 2018, p. 895–906.
55. Callejas Z, López-Cózar R, Ábalos N, et al. Affective conversational agents. In: *Conversational agents and natural language interaction*. Hershey, Pennsylvania: IGI Global, 2011, pp. 203–222.
56. Luger E and Sellen A. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 2016.
57. Mori M, MacDorman KF and Kageki N. The uncanny valley [from the field]. *IEEE Robot Autom Mag* 2012; 19: 98–100.
58. Seymour M, Yuan LI, Dennis A, et al. Have we crossed the uncanny valley? Understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems* 2021; 22: 9.
59. Ciechanowski L, Przegalinska A, Magnuski M, et al. In the shades of the uncanny valley: an experimental study of human–chatbot interaction. *Future Gener Comput Syst* 2019; 92: 539–548.
60. Ivanovic M and Semnic M. The role of agent technologies in personalized medicine. In: 2018 *5th International Conference on Systems and Informatics (ICSAI)* 2018, pp.299-304. IEEE.
61. Harms J-G, Kucherbaev P, Bozzon A, et al. Approaches for dialog management in conversational agents. *IEEE Internet Comput* 2018; 23: 13–22.
62. Nadarzynski T, Miles O, Cowie A, et al. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digit Health* 2019; 5: 2055207619871808. 2019/08/31.
63. Lyles CR, Sarkar U and Osborn CY. Getting a technology-based diabetes intervention ready for prime time: a review of usability testing studies. *Curr Diab Rep* 2014; 14: 34.

Appendix

Students’ conversation with Melinda according to usability characteristics of chatbots (effectiveness, efficiency and satisfaction)

Effectivity	Example
Task completion	**Conversation ends unanswered** Melinda: *Name*, would you say that your studies stress you out in general?
Accuracy	**Melinda does not understand the answers and continues to follow her algorithm** Melinda: I have a few questions about your studies. Do you have courses in presence? Person: Counter-question: What do I do this for? Melinda: Then you have certainly had online university anyway. Melinda: How did you feel about online-learning last semester, *name*? Person: *Name* must work now. **Melinda shares non-understanding** Melinda: Do you believe that you are being prepared well for the everyday life as a doctor, *name*? Person: Oh, I can’t tell that now, but that always depends on the initiative of each person. Melinda: Unfortunately, I did not understand this!
Efficiency	Example
Communication effort	**Questions from the Bot are unanswered** Melinda: Do you feel under time pressure? Melinda: Are you busy right now? Person: Sometimes **Short and brief answers of the students** Melinda: How are you doing, *name*? Person: well Melinda: That´s nice to hear! Melinda: Do you currently have a lot to do? Person: yes
Satisfaction	Example
Ease-of-use	**Irritated response from the students** Melinda: But do you have any hobbies? Person: Chess, running and weightlifting. Melinda: What is your hobby? Person: CHESS **No talkative attitude towards the Bot** Melinda: There is a lot going on right now. Do you want to tell me more about it *name*? Person: There is not much to talk about Melinda: Okay, then tell me. Person: Melinda: Yes, I understand that. I hope that soon you will be less burdened! Melinda: Do you feel energetic? Person: There is nothing to tell. **Subliminal criticism of Melinda** Melinda: And how are you currently doing with hybrid-teaching formats? Person: I find it very one-sided and miss the interaction with the other students Melinda: Which disadvantages do you see? Person: besides, I would find it much more pleasant to talk to a real person....

(continued)

Appendix. Continued.

Context-dependent questions	**Question–answer mechanism incorrect** Melinda: Do you sometimes have the feeling that the requirements are too high for you? Person: No, I'm not afraid of covid Person: No, I don't learn in a learning group **Melinda misunderstands answers** Melinda: Do you feel alone often? Person: No Melinda: That's too bad: (The pandemic is really hard….)
Learnability	**Repeated questioning does not lead to a learning effect** Melinda: What´s your name? :) Please just write your name. Person: *name* Melinda: So *name* is your name? Person: That's how it is. Melinda: Oh, did I get you wrong? Please write your name again now :) Person: *name* Melinda: So *name* is your name? Person: Eric Melinda: Oh, did I get you wrong? Please write your name again now :)