

Verifying Scientific Validity and Enhancing Response Motivation in Digital Human-Based Mental Health Assessment: A Mixed-Methods Study

Chi-Cheng Lao^{[0009-0008-6752-0269]¹}, Litun Tan², Jing Liang¹, Ling Zeng¹, Zongze Li², Yitian Huang², Zi-Yi Guo², Yan-Dian Xu², Tianyang Tan³, Xuetao Tian^{[0000-0001-8407-8718]¹(✉)}, Jun Liu²(✉), Keliang Chen²(✉)

¹Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, Beijing 100875, China
chichenglao@mail.bnu.edu.cn

²School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
chenkl@bupt.edu.cn, liujun@bupt.edu.cn

³School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China

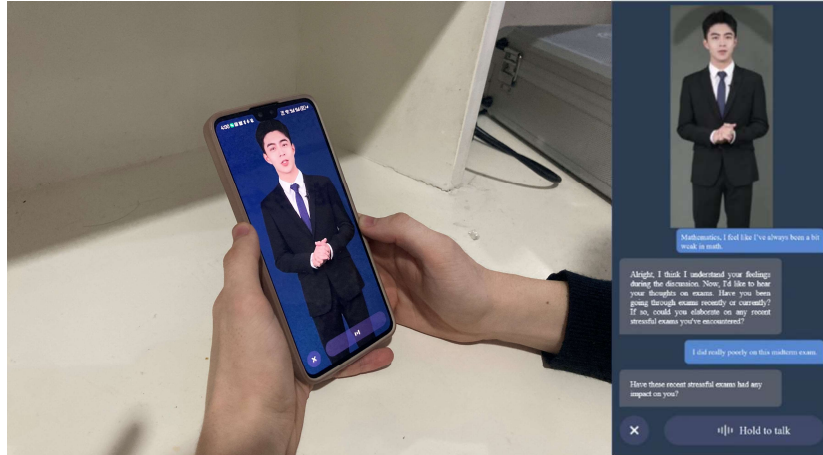


Fig. 1. On the left side of the image is a photo of people using the Digital Human-Based Interactive Mental Health Assessment System (DHMHAS), while on the right side is the user interface of DHMHAS.

Keliang Chen, Jun Liu, and Xuetao Tian are the first, second, and third corresponding authors, respectively.

Abstract. Online mental health surveys have long been challenged by respondent carelessness, a problem that has persisted for over a decade without a comprehensive solution. To tackle these issues, we designed and implemented the Digital Human-Based Interactive Mental Health Assessment System (DHMHAS) with college students as the pilot group, and employed mixed methods to validate its scientific credibility and interaction experience to comprehensively explore whether the system may retain scientific validity while increasing motivation to answer questions. We successfully confirmed that DHMHAS possesses robust psychometric properties. In particular, the system performed better than traditional methods in terms of criterion validity. Furthermore, DHMHAS enhanced the college students' interactive experience by providing a high level of social presence and system usability, reducing cognitive load, and eliciting high levels of response differentiation in answers. They also offered overwhelmingly positive feedback on the system, recognizing its benefits for personal growth and development. Compared to traditional methods, DHMHAS offered significant improvements in delivering positive emotional experiences, deepening cognitive understanding, promoting immersive engagement, and increasing the accuracy of responses. This research lays a scientific foundation for developing innovative assessment tools and offers directions for optimizing future AI-based mental health evaluation methods.

Keywords: careless response, digital human, human-AI interaction, user experience design

1 INTRODUCTION

The main source of mental health problems among college students is academic pressure [1, 2]. It has been an important issue in the current Chinese educational system [3, 4]. To solve these problems caused by academic stress, the first step is to evaluate college students with academic stress. Although traditional psychological assessment has been applied in wide-speed fields to assess the levels of mental health in individuals, the lack of motivation to answer questions (also called careless responses) is still a tricky challenge in recent decades [5–7]. This challenge could cause a series of issues in data quality, subsequently harming the actual levels of academic stress and mental health, and affecting the judgment and intervention of psychological counselors. To address these challenges, we designed and implemented a digital human-based interactive mental health assessment system (DHMHAS) targeted at college students. Due to the extensive anthropomorphism of digital humans in DHMHAS, college students can have a highly immersed interactive experience to improve the quality of their responses. Moreover, DHMHAS uses natural language to interact with individuals which can offer richer information [8, 9], enabling assessment outcomes to be more authentic and accurately reflect real-world contexts and behaviors.

Specifically, the digital human in DHMHAS would engage in conversations with students about academic stress questions using natural language and multi-turn dialogue. In the process of the interaction with DHMHAS and students, their answers will be processed through a semantic analysis and classification parts to evaluate their levels

of academic stress. Ultimately, an academic stress assessment report would be generated for psychological consultants to understand individual academic stress and potential mental health problems. Additionally, we validated and explored research questions during the development of the system to strengthen its scientific validity in psychological measurement, providing a scientific and effective academic stress management platform. The system implementation and validation are as follows in Fig.2.

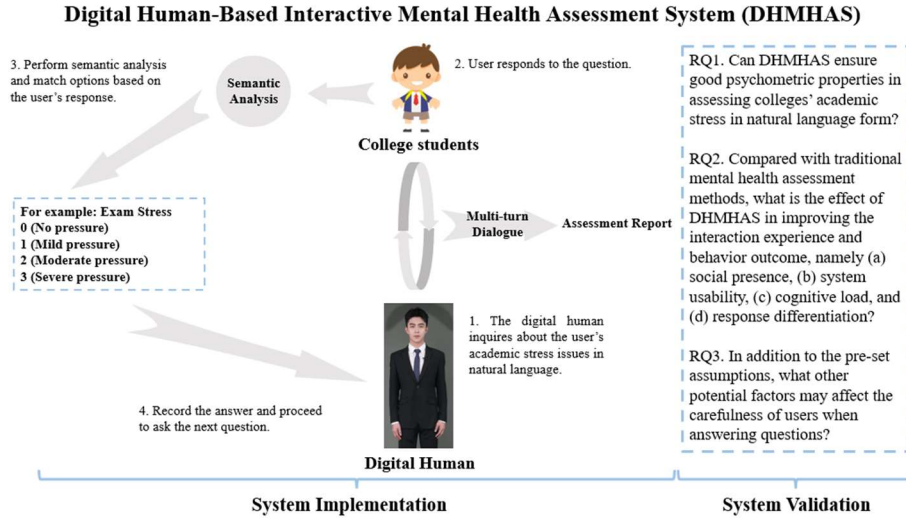


Fig. 2. The implementation and validation of DHMHAS.

2 RELATED WORK

2.1 Limitations of Traditional Mental Health Assessment

Due to the prevalence of the Internet, the most widely used method for mental health assessments at present is through online questionnaire responses [10]. Although this method allows for a rapid understanding of an individual's mental health status, it has a major limitation: Traditional mental health assessment methods are tedious and can easily reduce individuals' motivation to continue answering questionnaires, leading to careless responses. Careless responses from individuals can introduce errors in assessment results, consequently affecting the data quality [6, 11]. In other words, this can result in questionnaire data inaccurately reflecting an individual's mental health status. In such cases, parents, teachers, and mental health professionals may not be able to monitor and address group concerns, potentially leading to serious societal consequences.

Despite the issues being long identified, a universal solution on how to genuinely enhance individuals' motivation to answer questions seriously has yet to be found, spanning over a decade [5–7, 12]. Specifically, previous researchers often used warnings to alert individuals to answer carefully or directly provided external incentives [13,

14] to motivate individuals to respond. However, the latest research indicated that presenting warning instructions to participants is no longer effective in reducing individuals' careless responses [12]. Another study [7] attempted to use social exchange theory, offering participants discounts on food and beverages in exchange for careful responses, but the study results showed that providing rewards did not effectively reduce careless responses. In other words, the quest for finding methods to eradicate individuals' careless responses remains a long and challenging journey. Before delving into solutions, it is crucial to grasp the underlying reasons for low motivation to respond. This understanding enables us to explore effective strategies to enhance motivation for answering.

2.2 The Reasons for Low Response Motivation in Traditional Mental Health Assessment

Prior research has extensively examined the reasons for individuals' low motivation to respond when completing questionnaires, suggesting a variety of possible factors [13]. The most comprehensive study in this area was conducted by [11], who noted that questionnaire design (e.g., questionnaires that are too long), respondent characteristics (e.g., personality traits), and contextual factors (e.g., environmental distractions) may influence motivation to respond.

In terms of questionnaire design, researchers frequently encounter the challenge of striking a balance between the length of the questionnaire and the richness of the data collected. This trade-off can result in an excessive number of questions, leading to the phenomenon of "bubble hell" (i.e., presenting the questionnaire in the form of a large number of questions) [7], which often evokes negative emotions in the user when completing the questionnaire [15]. Consequently, in the absence of a positive interaction experience, it may be an unrealistic expectation to expect respondents to answer carefully.

In examining the characteristics of respondents, while some studies have indicated that individuals with diminished responsibility and self-control may exhibit less effort in questionnaire completion [16, 17], the correlation between personality traits (e.g., conscientiousness) and carefulness was not particularly strong ($r = .03 - .17$). Therefore, personality traits may not be the primary determinant of motivation to respond.

In terms of contextual factors, the absence of direct interaction [18] between the respondents and the researcher may be a potential cause of low individual motivation to respond. Humans have an innate need for social interaction [19], and the lack of interactivity may result in respondents not receiving immediate feedback, which may reduce their engagement and motivation to answer the questionnaire.

In conclusion, following the exclusion of personality factors, questionnaire design and contextual factors should be considered as potential areas for improvement, to increase users' motivation to respond. It is regrettable that, despite the undertaking of studies to suggest improvements in two factors to increase motivation to answer, the effectiveness of these attempts remains limited. For example, [20] attempted to enhance respondents' motivation to answer by establishing an avatar to simulate the researcher's supervision. However, this approach did not yield the anticipated outcomes. Potential

explanations for this phenomenon are a dearth of interactivity, an absence of realism in the avatar's image, and persisting deficiencies in the questionnaire design, which collectively contribute to an unsatisfactory interaction experience [21]. This, in turn, maybe a pivotal factor in reducing users' motivation to respond to the questionnaire.

2.3 Potential Innovative Methods for Enhancing Respondent Motivation

Regarding questionnaire design, by adopting a one-question-one-answer presentation, the cognitive load of the user can be effectively reduced, which in turn enhances the usability of the system. According to Cognitive Load Theory (CLT,[22]), intrinsic cognitive load (ICL) is related to the inherent complexity of the task and extraneous cognitive load (ECL) is related to the presentation of the task. In questionnaire answering, a design similar to "bubble hell", may affect the individual's perception of the difficulty of the task itself and the efficiency of information processing. Specifically, in terms of ICL, the inherent complexity of the task may lead individuals to perceive increased difficulty in carefully answering the questionnaire. On the ECL side, lengthy text requires individuals to sift through and process large amounts of information to find key elements to recall and fill in the answers. Both loads consume the user's limited cognitive resources [23]. Once these resources are depleted, individuals are less likely to focus on the task and exhibit low motivation to answer the questionnaire. Therefore, a new mental health assessment tool with a one-question-one-answer format can reduce the amount of information that users need to process at any given time, thereby reducing the ICL and ECL, resulting in a perceived high usability and a good response experience.

In terms of social interaction, it is crucial to enhance the anthropomorphism of avatars, which includes aspects such as their appearance, body movements and speech. Through a multimodal approach, i.e., combining visual, auditory, and other sensory experiences, it is possible for users to feel a higher level of positive experience when interacting with others, also known as social presence. Social Presence (SP) refers to the "degree of interaction with others and the significance of interpersonal relationships that follow," where these "others" can be real people or intelligent forms generated by computer-mediated communication (CMC), such as virtual human agents [24, 25]. In general, increasing the anthropomorphic features of virtual humans enables individuals to perceive a high level of SP [26, 27] because it provides humans with a rich sensory experience, allowing users to immerse themselves in the interaction and have a good interactive experience. Therefore, virtual humans in novel mental health assessment tools with highly anthropomorphic features enable users to perceive high levels of SP, and users are likely to have a more positive user experience when using the tool, and this experience also positively affects motivation for answering the questions and improves the accuracy of responses. Specially, the improvement of motivation in responses could be used [28] to calculate.

2.4 DHMHAS

To address the limitations of the traditional mental health assessment methods, we designed a novel system based on natural interaction for assessing mental health using a digital human (DHMHAS) to measure individual academic stress (the specific assessment process has been shown in “introduction”). The system mainly integrates (a) 3D digital human generation, (b) voice cloning, (c) semantic analysis, and (d) speech recognition technology, as shown in Fig. 3.

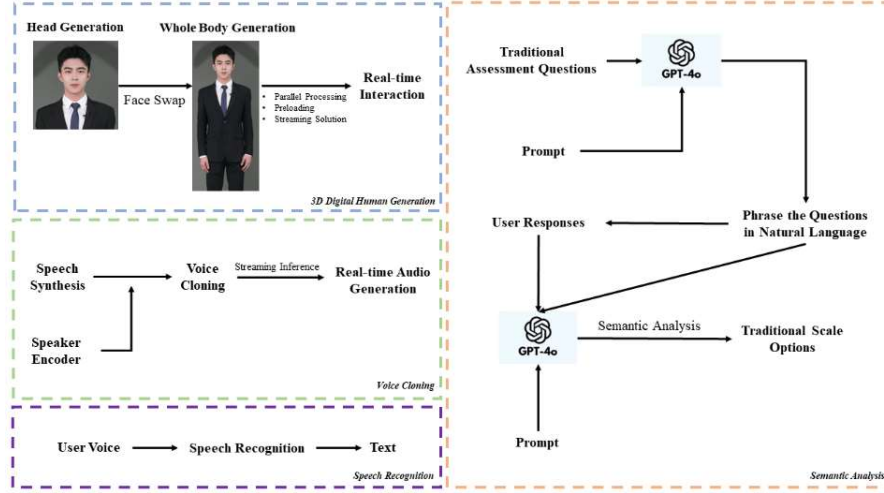


Fig. 3. Technical implementation diagram.

Based on the ER-NeRF [29] 3D digital human generation technology, we proposed a digital human face-changing solution to address the limitations of ER-NeRF that only generated the head, and the head was not dynamically scaled under control. Through the alignment algorithm, that is, fixing the non-face area and calculating the corresponding face position, the overall generation of the digital human head, torso, and limbs was achieved, thereby optimizing the appearance of the digital human. To solve the problem of the slow rendering speed of ER-NeRF and the slow generation speed of digital humans after the face-changing operation, which leads to the lag of digital humans, we maximized the use of GPU performance, preload, and selected advanced streaming solutions suitable RTMP [30] for this system to improve the generation efficiency, to meet the requirements of real-time interaction while ensuring high-quality image generation.

Voice cloning technology was used to generate audio with speaker timbre characteristics. Based on the VITS [31] speech synthesis model, we introduced a speaker encoder to characterize the speaker timbre characteristics and achieve the transformation from speech synthesis to voice cloning. To solve the problem of slow reasoning of long texts in the VITS model, we proposed a streaming reasoning solution, which receives text input segment by segment and generates speech in real-time without waiting for all text

input to be completed and sent to the digital human module, driving the real-time rendering and generation of the digital human.

Semantic analysis technology was mainly used to generate natural language questions about academic pressure and imitate the ability of humans to understand user intentions to match the options of traditional psychological assessment scales with the user's answers. Specifically, using GPT-4o, combined with a single-sample learning strategy and a psychological expert assessment strategy to construct a prompt project, traditional psychological assessment questions were converted into natural language expressions and output to users. Subsequently, using the semantic analysis capability of GPT-4o, the user's answers were processed to achieve a matching of options in the traditional psychological assessment scale.

Speech recognition technology converted the user's input speech into text for GPT-4o to analyze semantics. The speech recognition technology used in this system is FunASR [32]. FunASR is a basic speech recognition toolkit that provides convenient scripts and tutorials, supports inference and fine-tuning of pre-trained models, and has significant recognition effects on Mandarin Chinese.

By optimizing and integrating digital human appearance, rendering speed, and audio generation speed comprehensively, we aim to enhance the anthropomorphic qualities of digital humans to maximize the sense of social presence for users. In this process, users engage in a question-and-answer format, as well as employing natural language for cognitive processing, which serves to reduce the perceived cognitive load and enhance the overall interaction experience.

2.5 Scientific Validity in DHMHAS

Before exploring the prospective advantages of DHMHAS in bolstering response motivation, it is imperative to ascertain that the transition from a conventional questionnaire to a natural language questioning format does not introduce inaccuracies. Conversely, this is a point that scholars in non-psychology tend to overlook when developing systems. It is therefore essential to rigorously validate the scientific validity of the system. This process comprises the following key steps [33, 34]: (a) content validity analysis; (b) item analysis; (c) reliability analysis; (d) convergent validity; (e) discriminant validity; and (f) criterion-related validity. Collectively, these steps provide a comprehensive assessment of the scientific validity of DHMHAS.

Content validity evaluates the relevance and representativeness of a test's content, specifically its coverage of the intended subject matter, themes, or topics [33]. The assessment process is typically conducted by experts in the field of psychology, who validate content validity by calculating two key metrics: the item-level content validity index (I-CVI) and the scale-level content validity index (S-CVI) [35]. To assess the consistency of scores, the Kappa coefficient (K^*) [36] is calculated as a measure of consistency across raters. This ensures the reliability of the scores, thereby enhancing the accuracy and validity of the content validity assessment. In particular, I-CVI greater than or equal to .78 indicates good content validity. The evaluation criteria for K^* are: .40 to .59 is considered fair, .60 to .74 is deemed good, and greater than .74 is considered excellent.

Item analysis is a process of employing statistical techniques to examine the responses of subjects to test items to evaluate and screen the items themselves. It is typically conducted before evaluating the reliability and validity of a given test. Item-total correlation is a frequently employed statistical instrument in the context of item analysis. The item-total correlation coefficient measures the correlation between the score of a single item and the total score of the entire test. It is used to evaluate the contribution of the item to the overall test performance [34]. In general, the item-total correlation coefficient should fall within the range of .2 to .4, with values above .4 being preferable.

Reliability analyses are used to assess the consistency and stability of measurement instruments. Commonly used reliability coefficients are Cronbach's alpha, developed by [36], which measures the consistency of each question in a test to ensure the reliability of the test.

The concept of validity is pivotal in determining whether a test accurately measures the intended construct. Validity can be categorized into three sub-types: convergent validity, discriminant validity, and criterion-related validity. Convergent validity ensures that a test measures the same constructs as another test specifically designed for those constructs. In contrast, discriminant validity demonstrates that scores on a test are uncorrelated with scores from a test measuring an unrelated construct. Criterion-related validity evaluates how well test scores correlate with an external criterion, which could be actual behavior, achievements, or other recognized benchmarks. If users are asked questions in a natural language format by a digital human and receive scores assessing academic stress, these scores should exhibit a high correlation with traditional mental health assessments that measure the same academic stress for convergent validity. For discriminant validity, these scores should show weak correlations with constructs that measure completely different attributes, such as social presence. Regarding criterion-related validity, these scores should correlate to some extent with external standards, such as depression, which is often linked to academic stress [37, 38].

3 THE PRESENT STUDY

While DHMHAS presents the potential to address traditional mental health assessments, it is still necessary to address the following 3 research questions to solidify the system's scientific rigor and practicality:

RQ1. Can DHMHAS ensure good psychometric properties in assessing colleges' academic stress in natural language form?

RQ2. Compared with traditional mental health assessment methods, what is the effect of DHMHAS in improving the interaction experience and behavior outcome, namely (a) social presence, (b) system usability, (c) cognitive load, and (d) response differentiation?

RQ3. In addition to the pre-set assumptions, what other potential factors may affect the carefulness of users when answering questions?

We employed 2 studies to verify the scientific validity and explore the interactive experiences in DHMHAS. Study 1 consisted of two experiments, that addressed RQ1. In particular, experiment 1 aimed to confirm whether the natural language version of

academic stress can accurately measure the construct using an expert perspective (i.e., context validity). Experiment 2 used participants' data to verify the psychometric properties of DHMHAS to fully ensure the quality of the items (i.e., item analysis), reliability, and validity (including convergent validity, discriminant validity, and criterion-related validity).

Study 2 used mixed methods to explore the benefits of improving response motivation in DHMHAS compared to traditional mental health assessment, which addressed RQ2 and RQ3. First, we compared two methods (i.e., traditional mental health assessment and DHMHAS) on the four key interactive experience variables and behavior outcomes: social presence, system usability, cognitive load, and response differentiation. Second, we used qualitative interviews with DHMHAS users to supplement the evidence for improving response motivation in DHMHAS. In particular, we focused on the features that attracted users to the system, the usefulness of the system, and how the system affected the psychology of users, including emotions, cognitions, wills, and behaviors. Through these analyses, we aimed to gain a deeper understanding of the potential strengths of DHMHAS and demonstrated their value in improving the motivation of responses. It is hoped that a possible theoretical model can be constructed for this purpose.

4 METHOD

4.1 Participants

We recruited 5 psychological experts in experiment 1 of study 1. Specifically, the participants included 1 male doctoral student in clinical psychology, 2 male doctoral students in educational psychology, 1 female doctoral student in educational psychology, and 1 female university teacher in psychology. In experiment 2 of study 1, we recruited 39 college students (sophomore: 41.00%, junior: 33.33%, senior: 25.6%), comprising 15 male students (38.50%) and 24 female students (61.50%).

In Study 2, participants were recruited through a student pool from Beijing Normal University. A total of 79 college students aged between 18 and 25 were recruited. All participants volunteered to participate and signed informed consent before the experiment. Participants were randomly assigned to two experimental groups: In the DHMHAS group, 41 participants comprised college students (sophomore: 39.02%, junior: 34.15%, senior: 26.83%). 24 (58.54%) were female and 17 (41.46%) were male. In the Survey group, we recruited 38 participants who comprised college students (freshmen: 7.9%, sophomores: 34.2%, juniors: 44.7%, seniors: 13.2%). 11 (28.95%) were female and 27 (71.05%) were male.

4.2 Apparatus and Procedure

This study has been approved by the Beijing Normal University Ethics Committee (BNU202406210122). In experiment 1 of study 1, we employed WènjuànXíng (a type of survey platform; <https://www.wjx.cn/>) to collect the content validity evaluation by psychology experts. In particular, each page would display 4 elements: (a) a definition

of the “stress” concept, (b) a definition of the “academic stress” concept, (c) the original author’s question logic, and (d) the natural language version of academic stress. An example of evaluating content validity is “How relevant is this question (in natural language) to the original question?”. In experiment 2 of study 1, we employed the academic stress questions recognized by experts in DHMHAS and conducted a more comprehensive psychological measurement verification of the system. Before the experiment began, participants would be asked to go to a quiet, distraction-free space to experiment. Once it was confirmed that the participants were in suitable conditions for the experiment, they would be introduced to the usage of DHMHAS. When the participants received the response from the digital human “All interview has ended, we will provide you with feedback as soon as possible,” it signified the end of the assessment session. Following this, participants would need to respond to (a) the original scale of academic stress [39], (b) the SP scale [40], (c) PHQ-9 [41]. Upon completion of the answers, the experiment concluded. An example of a natural language version of an academic stress question could be shown in Fig.4.

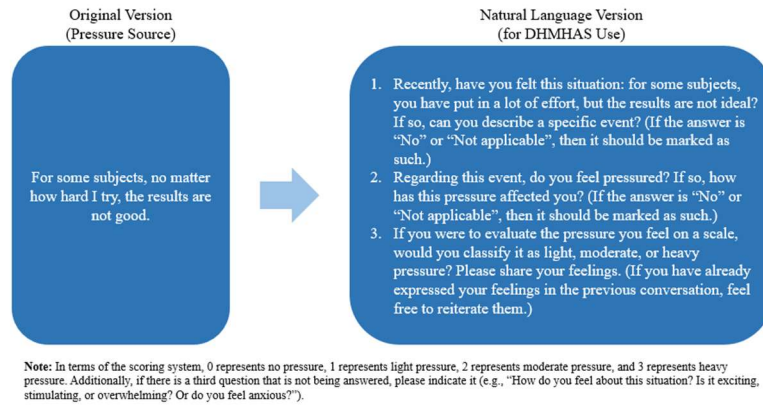


Fig. 4. An example of the natural language version of academic stress.

In Study 2, we recruited college students and randomly assigned them to two groups: the DHMHAS group and the Survey group. For the DHMHAS group, participants first signed an informed consent form and completed a pretest to collect demographic information and their experiences with artificial intelligence. They were instructed to participate in a quiet, distraction-free environment. Participants received an installation package for DHMHAS along with a user guide. They were briefed on the experiment’s purpose, app usage, and instructions before using the DHMHAS app. After the experiment, participants completed several questionnaires: the SP scale [40], the ICL and ECL scale (for measuring cognitive loads)[42], and the System Usability scale (for measuring the usability of DHMHAS) [43]. In the Survey group, participants used traditional psychological assessment methods to answer questions. Similar to the DHMHAS group, demographic information and familiarity with artificial intelligence were collected. Participants were asked to respond to questions about academic stress

in a quiet environment. After completing these tasks, they filled out the same questionnaires as the DHMHAS group. Finally, in terms of the qualitative interviews, participants in the DHMHAS group would undergo a brief interview focusing on their interactive experiences with DHMHAS, the questions are as follows:

- Are there any features or designs that impressed you?
- What help or impact does this product have had on you?
- In comparison to filling out traditional questionnaires, what advantages do you think this digital agent assessment method has?

4.3 Analysis

In experiment 1 of study 1, we conducted data analysis using content validity. In experiment 2 of Study 1, we employed the item-total correlation coefficient in item analysis. In terms of reliability, we utilized Cronbach’s alpha (α) to evaluate the consistency of academic stress scores obtained by digital humans through questions asked in natural language form. In terms of convergent validity, discriminant validity, and criterion-related validity, we assessed the validity of our measures by calculating Pearson correlations between the natural language version and the original version of academic stress, SP (for discriminant validity), and depression (for criterion-related validity), respectively. In Study 2, we utilized an independent t-test to explore the differences in interactive experiences (i.e., SP, ICL, ECL, system usability, and response differentiation [28]) between the DHMHAS group and the Survey Group. Finally, a Grounded Theory [44] would be used with first, second, and third levels for qualitative interview coding.

In terms of analysis software, we conducted data analysis using SPSS 27.0 in Study 1 and Study 2. Excel was employed for interview coding in Study 2.

5 RESULTS

5.1 Study 1

Content validity. Five experts conducted a content validity assessment on natural language items regarding academic stress. All items met the content validity criteria, with an Item-Content Validity Index (I-CVI) greater than .78 for all questions, and a Kappa coefficient exceeding .74.

Psychological measurement item analysis. The correlations between each item and the total score range from .40 to .67, all greater than or equal to .4.

Reliability. Regarding internal consistency reliability measured by Cronbach’s alpha, the natural language version of academic stress items demonstrated good reliability ($\alpha = .77$).

Convergent validity, discriminant validity, and criterion-related validity. Regarding convergent and discriminant validity, the natural language version of academic stress exhibited a high linear correlation with the original version of the academic stress ($r = .82, p < .001$) and a low linear correlation with SP ($r = .28, p = .10$), respectively. In terms of criterion-related validity, the natural language version of academic stress was significantly correlated with depression ($r = .62, p < .001$). Specifically, the correlation between the natural language version of academic stress and depression was higher than that between the original version of academic stress and depression ($r = .56, p < .001$), as shown in Table. 1.

Table 1. Convergent Validity, Discriminant Validity, and Criterion-related Validity.

	<i>M</i>	<i>SD</i>	1	2	3
1 Academic Stress (Natural Language Version)	13.06	5.38	-		
2 Academic Stress (Original Version)	14.44	5.28	.82***	-	
3 Social Presence (SP)	48.72	9.62	.28	.29	-
4 Depression	5.81	4.58	.62***	.56***	-.09

Note. $N = 36$, * $p < .05$, ** $p < .01$, *** $p < .001$

5.2 Study 2

Differences in interactive experience and careless behavior

Social presence. A significant difference was observed in participants' SP ($p = .02, t = -2.30$, Cohen's $d = 1.05$) between the Survey group ($M = 6.75, SD = 1.20$) and the DHMHAS group ($M = 7.23, SD = .89$).

Cognitive load. Compared to the Survey group ($M = 4.74, SD = 1.30$), the DHMHAS group ($M = 2.89, SD = 1.30$) had significantly lower ICL ($p < .001, t = 6.30$, Cohen's $d = 1.30$) and ECL ($p < .001, t = 8.21$, Cohen's $d = 1.36$).

System usability. The DHMHAS group ($M = 78.29, SD = 10.22$) demonstrated higher system usability compared to the Survey group ($M = 67.04, SD = 19.16$), and this difference was significant ($p = .002, t = 3.29$, Cohen's $d = 15.19$).

Carefulness behavior. The DHMHAS group ($M = .69, SD = .19$) demonstrated higher response differentiation compared to the Survey group ($M = .56, SD = .08$), and this difference was significant ($p < .001, t = 3.71$, Cohen's $d = .15$).

Interview results 1: Impressive features and designs in digital humans.

Digital Human emerged as the most mentioned, capturing the majority of participants' attention towards impressive features or designs, reaching 73.08%. Among these, the

Appearance (31.58%), Speech (42.11%), Body Language (13.16%), and Facial Expressions (13.16%) of the digital human were identified as the most noteworthy aspects.

- Appearance: “...*this digital human was quite real.*” (Participant #27)
- Speech: “*The character above would exhibit certain actions, along with speech, to make you feel like you were experiencing interaction with him.*” (Participant #1)
- Body Language: “*I thought his action was quite impressive. When you were speaking, its hands moved, giving the feeling that it resembled a real person, making you feel like the other party was seriously listening to your narration as you responded. Additionally, it nods and such, making you feel that what you were saying was being acknowledged and heard.*” (Participant #34)
- Facial Expressions: “...*when he answered me, his mouth would also move, which surprised me.*” (Participant #19)

Interview results 2: System usefulness.

The majority of responses revolved around the topic of *Personal Growth and Development*, accounting for 90.91% of the total mentions. There were some examples of second-level and third-level coding:

- Emotional Management (20.00%): “*After the conversation just now, I felt that some analysis was done on my recent state, which also led me to reflect on some aspects of myself.*” (Assisting in reflecting on one’s emotions or state; Participant #1)
- Emotional Support (40.00%): “*I felt that I could express my thoughts more effectively. Some things that I couldn’t express to friends or family around me, I could share with that artificial intelligence to seek help.*” (Accompanying listening; Participant #2)
- Self-Exploration (32.50%): “*By having it go over things with you, you could understand your shortcomings and what troubled you. You could speak them out to it. This helped you see your deficiencies and also recognize your strengths.*” (Understanding oneself; Participant #33)
- Problem Solving (7.50%): “*I felt that this product’s features were quite good; it could help me solve some issues in my daily life that I couldn’t handle on my own. For example, when it came to my emotions or when I felt mentally uncomfortable, I didn’t feel like I was alone. I didn’t feel as lonely, and I could consult it for advice, that was the approach I took.*” (Resolving issues; Participant #19)

Interview result 3: The advantages of DHMHAS based on psychology variables.

The vast majority of responses (94.55%) were focused on key psychological aspects: Emotions (32.73%), Cognitions (21.82%), Wills (29.09%), and Behaviors (10.91%).

Regarding Emotions, participants perceived that DHMHAS could help in avoiding boredom, tedium, and fatigue (50.00%), experiencing novelty (11.11%), having fun (22.22%), feeling friendly and affable due to the digital human’s responses (11.11%), and feeling be empathized (5.56%):

- Avoiding boredom, tedium, and fatigue: “...It won't be as dull. It won't be as boring as the traditional way of filling out questionnaires.” (Participant #18)
- Experiencing novelty: “One of its advantages is that it is innovative and more easily accepted by young people.” (Participant #3)
- Having fun: “I think it might be more interesting than traditional questionnaires. And if you simply fill out a questionnaire, you might find it a bit dull and boring.” (Participant #24)
- Feeling friendly and affable: “...Not as rigid as a questionnaire, it has a tone that feels like a psychological counselor. This may make you feel quite close to it, allowing you to share any issues you might have. You might feel like confiding in it and seeking advice or solutions from it.” (Participant #10)
- Feeling empathized: “The questionnaire format fails to grasp my feelings adequately, seeming to lack the empathetic emotion with me, making the questionnaire format too monotonous.” (Participant #14)

Concerning *Cognitions*, participants believed that DHMHAS could facilitate *easy and deep problem understanding* (66.67%), *deep self-understanding* (25.00%), and *easier retrieval of key information* (8.33%):

- Easy and deep questions understanding: “The advantage was the ability to delve deeper into understanding (questions). It added pressure and categorizes it into mild, moderate, and severe levels. It also followed up with questions about the reasons behind these aspects. In contrast, questionnaires were superficial.” (Participant #43)
- Deep self-understanding: “Sometimes he asked me what aspects or areas of study I felt stressed about, and I might have found subjects like physics more challenging, especially science subjects. Sometimes I also reflected on myself and felt that I might not have been good enough in some aspects. I procrastinated a bit when doing homework, and occasionally I felt a little regretful, thinking that I could do better and needed to make some improvements.” (Participant #10)
- Easier retrieving of key information: “...I felt that conducting surveys in this manner was beneficial. He could read it out to me, and I thought this was quite helpful in quickly grasping the important information within the instructions” (Participant #35)

Regarding *Wills*, participants found that this interactive approach could enhance focus, engagement, and concentration (62.50%), elicit relaxed responses (18.75%), and boost motivation to answer questions (18.75%):

- Enhancing focus, engagement, and concentration: “Sometimes, when filling out surveys, details could be overlooked. Digital human narration could capture your attention and help you better understand the questions!” (Participant #33)
- Eliciting relaxed responses: “...His expressions and his movements made you feel like you were interacting with a... anyhow, not like dealing with a machine or a questionnaire. In short, his expressions, his movements, that kind of thing. They

would make you feel quite relaxed as if you were communicating with a real human professor.”

- Boosting motivation to answer questions: “The advantage was... Honestly, the person on the other end made me more willing to respond a bit because they gave off a dynamic vibe. That format involved a person with body language and some other elements. I wasn’t sure if there were facial expressions, but it felt like there might have been a hint of expression. It truly was because... Compared to questionnaires, it certainly was more likely to stimulate the idea of participants responding in this way, I believed.” (Participant #22)

On the topic of Behaviors, participants noted that this could lead to more accurate and comprehensive responses (100.00%). For example: “I felt that in this assessment method (DHMHAS), my responses would be more authentic.” (Participant #4)

6 DISCUSSION

6.1 Scientific validity of DHMHAS

The experiment in study 1 provides preliminary evidence of the content validity of the DHMHAS in psychological measurement, demonstrating that DHMHAS aligns with psychometric properties. Results indicated that the natural language version of academic stress could measure the original construct [35] in college students’ groups.

In study 2, building upon the findings of study 1, a further analysis of the scientific rigor of the natural language version of academic stress was conducted from theoretical perspectives. Results suggested that most evidence points towards the psychological scientific validity of the system. Specifically, (a) item analysis: item-total correlations ranged from .40 to .67; (b) reliability: the natural language version of academic stress demonstrated good reliability ($\alpha = 0.77$); (c) convergent validity: high correlation between the original version and the natural language version of academic stress; (d) discriminant validity: low correlation between the natural language version of academic stress and social presence; (e) criterion validity: significant and moderate correlation between the natural language version of academic stress and depression, consistent with prior research [38].

A new exploration is that the study found that the criterion-related validity of the natural language version of academic stress with depression was higher compared to the original version. It may indicate that this measurement tool is more effective in assessing individual academic stress. This is because natural language allows individuals to think about and respond to questions from an ecological perspective [8, 9], leading to more accurate measurements and a closer association with the negative emotions often associated with stress, such as depression. In summary, for RQ1, DHMHAS may be superior in terms of accuracy and reliability. The system is not only scientifically sound but also demonstrates better criterion-related validity.

6.2 Improvement in motivation to respond

Study 2, building on the established scientific validity of the DHMHAS system, delves deeper into its potential to enhance the sincerity of student responses. The effectiveness of DHMHAS is verified through comparative analysis of interaction experiences and response variability indices, supplemented by qualitative interviews to address research question 2. The findings indicate that DHMHAS significantly improves the level of earnestness in user responses.

Compared to traditional methods, DHMHAS offered a higher degree of SP due to its natural language interaction capabilities. Qualitative interview results revealed that students were notably impressed with digital humans and experienced a sense of engaging communication, which was stimulated by the highly anthropomorphized digital humans. Additionally, the perceived social presence elicited by DHMHAS led to lower intrinsic and extrinsic cognitive loads during interactions. This reduced cognitive load may lead them to perceive DHMHAS as more user-friendly, thereby enhancing their willingness to respond actively, such as the focused and relaxed responses mentioned in interviews. Behaviorally, this willingness to respond actively was reflected in higher response variability, which is not the result of random responses but rather a thoughtful deliberation by users to respond more accurately and comprehensively, fitting the real-life scenarios.

Furthermore, in the process of measuring academic stress, we found that the reduction in cognitive load was not merely achievable through interface design but required the integration of social interaction elements. According to Maslow's hierarchy of needs, humans have fundamental social needs [45]. When these needs are met, positive experiences emerge. Thus, when students answered surveys through the DHMHAS, the system engaged in dialogue using natural language with highly anthropomorphized digital humans, making them feel as though they were interacting with real people [26, 27]. Since humans are inherently skilled at social interactions, this mode of communication reduces the perceived task difficulty, thereby decreasing the internal cognitive load compared to traditional methods of answering psychological health surveys. Regarding external cognitive load, users can directly contemplate and recall in response to questions posed by digital humans, without the need to search for key information within lengthy surveys, significantly lightening their cognitive burden. Therefore, this digital-human interaction not only satisfies users' social needs but also optimizes their experience in answering psychological health surveys, making the task feel more relaxed and natural.

Given these potential advantages, we recommend considering the design of a truly human-centered survey response product in future practical activities. Such an interaction form can enhance users' willingness to respond earnestly, thus improving the user experience. Additionally, we hope to continue exploring the boundaries of this survey response method. For instance, research is needed to determine for which populations this interactive method can enhance the answering experience and for which populations traditional survey methods might be more appropriate. We also need to explore the optimal level of anthropomorphism to enhance the answering experience, providing more possibilities for future development.

6.3 Other factors influencing survey quality and proposed theoretical models.

In addition to the variables we anticipated affecting individual response experiences, interview results have also revealed the advantages of DHMHAS in terms of perceived technological usefulness and other psychological variables, such as positive emotional experiences (e.g., avoiding boredom) and a deeper cognitive understanding of questions. These factors could also be key in influencing whether users are willing to continue responding earnestly.

Based on the Technology Acceptance Model (TAM) [46], when a technology is perceived as useful by an individual, it enhances their positive attitude toward the system, thereby increasing the likelihood of them engaging in positive behaviors. From the current interview results, users could better understand their academic issues and reflect on areas needing improvement after interacting with the system, indicating that they found the system helpful and were thus more willing to engage in the dialogue process. In terms of emotional variables, DHMHAS addresses the dullness associated with traditional psychological assessments [15]. In cognitive terms, the questioning style of DHMHAS has ecological validity, prompting users to provide more detailed and accurate responses [8, 9].

Integrating these quantitative and interview findings, we have constructed a potential theoretical model based on the Theory of Planned Behavior [47] and the TAM [46]. This model explores how to enhance users' positive experiences and strengthen their motivation to respond through highly anthropomorphized means such as digital humans, thereby improving data quality. This model primarily involves three core variables: antecedent variables are individual perceptions of the system (e.g., anthropomorphism and social presence), potential mediating variables are psychological factors, including emotions, cognition, and intentions, which are significant research variables that may interact and influence each other, ultimately affecting the user's actual behavior—whether they are willing to respond earnestly to surveys, as shown in the response variability in Fig. 3.

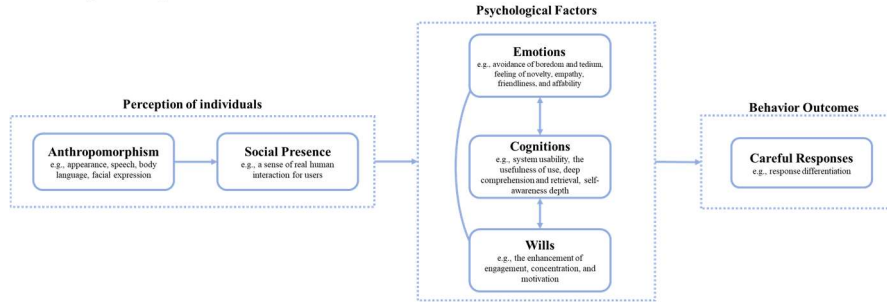


Fig. 5. A theoretical framework for preventing careless responses.

7 ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grant 62306044, in part by National Social Science and Technology Fund

(BGA230254). To enhance the readability of the text, we utilized GPT-4 for grammar checking and refinement.

8 REFERENCES

1. Karatekin, C.: Adverse Childhood Experiences (ACEs), Stress and Mental Health in College Students. *Stress Health*. 34, 36–45 (2018). <https://doi.org/10.1002/smi.2761>.
2. Pedrelli, P., Nyer, M., Yeung, A., Zulauf, C., Wilens, T.: College Students: Mental Health Problems and Treatment Considerations. *Acad. Psychiatry*. 39, 503–511 (2015). <https://doi.org/10.1007/s40596-014-0205-9>.
3. Wu, D., Yu, L., Yang, T., Cottrell, R., Peng, S., Guo, W., Jiang, S.: The Impacts of Uncertainty Stress on Mental Disorders of Chinese College Students: Evidence From a Nationwide Study. *Front. Psychol.* 11, (2020). <https://doi.org/10.3389/fpsyg.2020.00243>.
4. Zhang, J., Zheng, Y.: How do academic stress and leisure activities influence college students' emotional well-being? A daily diary investigation. *J. Adolesc.* 60, 114–118 (2017). <https://doi.org/10.1016/j.adolescence.2017.08.003>.
5. Francavilla, N.M., Meade, A.W., Young, A.L.: Social Interaction and Internet-Based Surveys: Examining the Effects of Virtual and In-Person Proctors on Careless Response. *Appl. Psychol.* 68, 223–249 (2019). <https://doi.org/10.1111/apps.12159>.
6. Meade, A.W., Craig, S.B.: Identifying careless responses in survey data. *Psychol. Methods*. 17, 437–455 (2012). <https://doi.org/10.1037/a0028085>.
7. Ward, M.K., Meade, A.W.: Applying social psychology to prevent careless responding during online surveys. *Appl. Psychol. Int. Rev.* 67, 231–263 (2018). <https://doi.org/10.1111/apps.12118>.
8. Kjell, O.N.E., Kjell, K., Garcia, D., Sikström, S.: Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods*. 24, 92–115 (2019). <https://doi.org/10.1037/met000191>.
9. Kjell, O.N.E., Sikström, S., Kjell, K., Schwartz, H.A.: Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Sci. Rep.* 12, 3918 (2022). <https://doi.org/10.1038/s41598-022-07520-w>.
10. Tao, Y., Zou, X., Tang, Q., Hou, W., Wang, S., Ma, Z., Liu, G., Liu, X.: Mapping network connection and direction between anxiety and depression symptoms across the early, middle, and late adolescents: Insights from a large Chinese sample. *J. Psychiatr. Res.* 169, 174–183 (2024). <https://doi.org/10.1016/j.jpsychires.2023.11.035>.
11. Ward, M.K., Meade, A.W.: Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices. *Annu. Rev. Psychol.* 74, 577–596 (2023). <https://doi.org/10.1146/annurev-psych-040422-045007>.
12. Brühlmann, F., Memeti, Z., Aeschbach, L.F., Perrig, S.A.C., Opwis, K.: The effectiveness of warning statements in reducing careless responding in crowdsourced online surveys. *Behav. Res. Methods*. 56, 5862–5875 (2024). <https://doi.org/10.3758/s13428-023-02321-z>.
13. Gibson, A.M., Bowling, N.A.: The Effects of Questionnaire Length and Behavioral Consequences on Careless Responding. *Eur. J. Psychol. Assess.* 36, 410–420 (2020). <https://doi.org/10.1027/1015-5759/a000526>.

14. Toich, M.J., Schutt, E., Fisher, D.M.: Do you get what you pay for? Preventing insufficient effort responding in MTurk and student samples. *Appl. Psychol.* 71, 640–661 (2022). <https://doi.org/10.1111/apps.12344>.
15. Ashley, M., Shaughnessy, K.: Predicting insufficient effort responding: The relation between negative thoughts, emotions, and online survey responses. *Can. J. Behav. Sci. Rev. Can. Sci. Comport.* (2021).
16. Barber, L.K., Barnes, C.M., Carlson, K.: Random and Systematic Error Effects of Insomnia on Survey Behavior. *Acad. Manag. Proc.* 2013, 11338 (2013). <https://doi.org/10.5465/ambpp.2013.11338abstract>.
17. Ward, M.K., Meade, A.W., Allred, C.M., Pappalardo, G., Stoughton, J.W.: Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Comput. Hum. Behav.* 76, 417–430 (2017). <https://doi.org/10.1016/j.chb.2017.06.032>.
18. Johnson, J.A.: Ascertaining the validity of individual protocols from Web-based personality inventories. *J. Res. Personal.* 39, 103–129 (2005). <https://doi.org/10.1016/j.jrp.2004.09.009>.
19. Lieberman, M.D.: *Social: Why our brains are wired to connect*. Broadway Books (2013).
20. Ward, M.K., Pond III, S.B.: Using virtual presence and survey instructions to minimize careless responding on internet-based surveys. *Comput. Hum. Behav.* 48, 554–568 (2015). <https://doi.org/10.1016/j.chb.2015.01.070>.
21. Tsai, W.-H.S., Liu, Y., Chuan, C.-H.: How chatbots' social presence communication enhances consumer engagement: the mediating role of parasocial interaction and dialogue. *J. Res. Interact. Mark.* 15, 460–482 (2021). <https://doi.org/10.1108/JRIM-12-2019-0200>.
22. Sweller, J.: Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* 12, 257–285 (1988). [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7).
23. Kahneman, D.: *Attention and effort*. prentice-Hall (1973).
24. Biocca, F., Harms, C., Burgoon, J.K.: Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence Teleoperators Virtual Environ.* 12, 456–480 (2003). <https://doi.org/10.1162/105474603322761270>.
25. Gefen, D., Straub, D.: Managing User Trust in B2C e-Services. *E-Serv. J. - E Serv J.* 2, 7–24 (2003). <https://doi.org/10.2979/ESJ.2003.2.2.7>.
26. Han, M.C.: The Impact of Anthropomorphism on Consumers' Purchase Decision in Chatbot Commerce. *J. Internet Commer.* 20, 46–65 (2021). <https://doi.org/10.1080/15332861.2020.1863022>.
27. Konya-Baumbach, E., Biller, M., von Janda, S.: Someone out there? A study on the social presence of anthropomorphized chatbots. *Comput. Hum. Behav.* 139, 107513 (2023). <https://doi.org/10.1016/j.chb.2022.107513>.
28. Heerwegh, D., Loosveldt, G.: Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality. *Public Opin. Q.* 72, 836–846 (2008). <https://doi.org/10.1093/poq/nfn045>.
29. Wang, X., Ruan, T., Xu, J., Guo, X., Li, J., Yan, F., Zhao, G., Wang, C.: Expression-aware neural radiance fields for high-fidelity talking portrait synthesis. *Image Vis. Comput.* 147, 105075 (2024). <https://doi.org/10.1016/j.imavis.2024.105075>.
30. A. Nurrohman, M. Abdurrohman: High Performance Streaming Based on H264 and Real Time Messaging Protocol (RTMP). In: 2018 6th International Conference on Information

- and Communication Technology (ICoICT). pp. 174–177 (2018). <https://doi.org/10.1109/ICoICT.2018.8528770>.
31. Kim, J., Kong, J., Son, J.: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, <https://arxiv.org/abs/2106.06103>, (2021).
 32. Sun, W.-R., Li, J.-H., Liu, L., Kevrekidis, P.G.: The instabilities beyond modulational type in a repulsive Bose-Einstein condensate with a periodic potential, <https://arxiv.org/abs/2305.00133>, (2023).
 33. Cohen, R.J., Swerdlik, M.E., Phillips, S.M.: Psychological testing and assessment: An introduction to tests and measurement. Mayfield Publishing Co (1996).
 34. Miller, L.A., Lovler, R.L.: Foundations of psychological testing: A practical approach. Sage publications (2018).
 35. Lynn, M.R.: Determination and quantification of content validity. *Nurs. Res.* 35, 382–385 (1986). <https://doi.org/10.1097/00006199-198611000-00017>.
 36. Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46 (1960). <https://doi.org/10.1177/001316446002000104>.
 37. Jayanthi, P., Thirunavukarasu, M., Rajkumar, R.: Academic stress and depression among adolescents: A cross-sectional study. *Indian Pediatr.* 52, 217–219 (2015). <https://doi.org/10.1007/s13312-015-0609-y>.
 38. Jiang, S., Ren, Q., Jiang, C., Wang, L.: Academic stress and depression of Chinese adolescents in junior high schools: Moderated mediation model of school burnout and self-esteem. *J. Affect. Disord.* 295, 384–389 (2021). <https://doi.org/10.1016/j.jad.2021.08.085>.
 39. LI Hong, K.W.-B.: Development of Stress Scale for College student. *Chin. J. Appl. Psychol.* 27–32 (2002).
 40. Lee, K.M., Peng, W., Jin, S.-A., Yan, C.: Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. *J. Commun.* 56, 754–772 (2006). <https://doi.org/10.1111/j.1460-2466.2006.00318.x>.
 41. Kroenke Kurt, Spitzer Robert L: The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatr. Ann.* 32, 509–515 (2002). <https://doi.org/10.3928/0048-5713-20020901-06>.
 42. Klepsch, M., Schmitz, F., Seufert, T.: Cognitive Load Measurement--Naïve Rating Questionnaire, <http://doi.apa.org/getdoi.cfm?doi=10.1037/t70903-000>, (2019). <https://doi.org/10.1037/t70903-000>.
 43. Brooke, J.: SUS: A “Quick and Dirty” Usability Scale. Presented at the (1996).
 44. Glaser, B., Strauss, A.: Discovery of grounded theory: Strategies for qualitative research. Routledge (2017).
 45. Maslow, A.H.: A theory of human motivation. *Psychol. Rev.* 50, 370–396 (1943). <https://doi.org/10.1037/h0054346>.
 46. Davis, F.D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* 13, 319–340 (1989). <https://doi.org/10.2307/249008>.
 47. Ajzen, I.: The Theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* (1991).