Educational technology research and development
https://doi.org/10.1007/s11423-026-10634-x

RESEARCH ARTICLE

Exploring student anxiety and experience in performance-based assessments using AIvaluate: an LLM-augmented emotionally intelligent pedagogical AI conversational agent

Habeeb Yusuf[1] · Arthur Money[1] · Damon Daylamani-Zad[1]

Received: 29 January 2025 / Accepted: 6 May 2026

Abstract
Performance-based assessments such as oral presentations and viva voce exams are valued for their pedagogical benefits but can also be associated with heightened student anxiety, which may affect a range of learners and hinder student motivation and overall assessment experience. This study explores the potential of AIvaluate; an emotionally intelligent, LLM-augmented conversational agent, to provide a more supportive assessment environment for such learners. Using a counterbalanced quasi-experimental, within-subjects design with 35 pre-university students, we compared experiences of traditional face-to-face assessments with AIvaluate-mediated sessions. Emotional state reports, usability ratings and qualitative feedback were analysed to evaluate anxiety, system usability and learner perceptions. Results indicated that students experienced significantly lower anxiety during AIvaluate sessions, and usability scores for the tool were rated in the "good" range. Thematic analysis further revealed perceived advantages such as reduced social pressure, flexible pacing and ease of use, alongside limitations including technical challenges and a lack of dynamic human interaction. Importantly, while these findings suggest that AI-mediated PBAs may reduce student anxiety, which may be beneficial for learners who experience heightened levels of anxiety that exceeds the individual zone of optimal functioning, further investigation is necessary to ascertain whether this is ultimately beneficial for learning and whether AI-mediated assessments can improve student attainment. Future research could explore the value of hybrid models that combine AI conversational agent-based and face-to-face assessment formats, and examine the long-term effects of AI-led assessments on student performance and well-being.

Keywords Artificial intelligence · Conversational agent · Chatbot · Assessment · Education · Emotional Intelligence · Performance based assessment · Student anxiety · Student well-being · Generative AI · LLM

Extended author information available on the last page of the article

Springer

Background

Performance-based assessments in education

Performance-based assessments (PBAs) are a type of pedagogical assessment method that requires students to demonstrate their knowledge, skills, and abilities by performing tasks or producing work whilst being assessed in real-time (Darling-Hammond, 1994; Doug et al., 1988). PBAs can include oral presentations, debates, exhibitions, performances, viva voce exams, and technical or visual demonstrations. PBAs are considered to be an effective pedagogical assessment method which can be used to determine the extent of retainment and understanding of students' knowledge in a particular subject, thus identifying whether learning has occurred, and to what degree. According to Baker (1997), PBAs focus more on assessing the students' cognitive abilities that synthesise and demonstrate comprehension across various disciplines, as opposed to traditional examinations, which tend to focus more on fostering retention of factual knowledge through memorisation. The use of PBAs within programmes of study can cultivate an interdisciplinary understanding of concepts in students, fostering their ability to discern interconnectedness and apply multifaceted knowledge in complex problem-solving scenarios. Therefore, PBAs are not only used in arts and humanities based programmes, but are also contemporarily used within science, technology, engineering and maths (STEM) programmes as they can more accurately predict a student's ability to apply learnt knowledge (Ernst, 2008; Potter et al., 2017). Due to the many advantages over traditional summative examinations, countries across Europe, Asia, Australia, and Latin America are restructuring their educational systems to incorporate PBAs across various levels, including primary, secondary, vocational, and tertiary education (Gallardo, 2020). The implementation of PBAs, however, creates challenges for the student experience, most notably student anxiety; these issues are explored in detail in the subsequent sections.

Performance-based assessments in the age of generative AI

The rise of generative AI has brought unprecedented advancements in technology, enabling sophisticated text and multimedia generation capabilities that have transformed the educational landscape (Eke, 2023; Tzirides et al., 2023). However, this technological progress has also introduced significant challenges, particularly in ensuring the integrity and authenticity of assessments. Traditional coursework submissions are increasingly susceptible to questions of plagiarism and authorship, as generative AI applications and large language models can produce high-quality outputs that mimic human writing, making it difficult to determine the originality of a student's work (Alser & Waisberg, 2023; Cotton et al., 2023; Dwivedi et al., 2023; Eke, 2023). PBAs have emerged as a critical response to these challenges, offering an assessment approach that inherently prioritises authenticity and integrity (Cotton et al., 2023; Pearce & Chiavaroli, 2023; Souppez et al., 2023). By evaluating students' skills and knowledge through active real-time demonstration, whether in the form of oral examinations, practical tasks, or hands-on problem-solving, PBAs provide a robust framework for mitigating the risks associated with generative AI-assisted plagiarism. Unlike traditional assessments, which primarily focus on written submissions or memory recall, PBAs requires students to engage dynamically with their examiners or the assessment process, thereby ensuring the genuineness of their contributions (Kabbar & Barmada, 2024; Newell, 2023). As PBAs have the potential of safeguarding the credibility of academic qualifications, they are a pivotal tool for educators in upholding academic standards in an era of rapid technological evolution.

Student anxiety

A discourse on student anxiety

Student anxiety is frequently positioned within educational research as an inherently negative affective state that undermines learning, assessment performance, and wellbeing. This position is supported by a substantial body of evidence linking heightened anxiety to attentional disruption, intrusive worry, and reduced working memory capacity (Angelidis et al., 2019; Eysenck, 2012). However, conceptualising anxiety as uniformly detrimental risks oversimplifying a construct that is more accurately understood as multidimensional and context dependent. Contemporary research increasingly emphasises that anxiety varies in both intensity and function, and that its effects on performance depend on where it is situated along a continuum of arousal (Chin et al., 2017; Weinberg & Mellano, 2023). A foundational framework for understanding this relationship is the *Yerkes–Dodson Law*, often described as an inverted U-shaped relationship between arousal and performance (Yerkes & Dodson, 1908). According to this model, low levels of arousal are associated with under-engagement and reduced motivation, whereas moderate levels of anxiety can be facilitative, supporting alertness, attentional focus, and task engagement. Beyond this optimal zone, further increases in anxiety lead to sharp declines in performance as cognitive resources become overloaded (Arent & Landers, 2003; Eysenck, 2012). Empirical work has demonstrated that moderate arousal is associated with optimal performance, while excessive anxiety impairs both accuracy and response efficiency (Arent & Landers, 2003).

Importantly, this arousal-performance relationship is not uniform across individuals. Models such as the *Individual Zones of Optimal Functioning (IZOF)* highlight that optimal anxiety levels vary considerably between learners, shaped by factors including trait anxiety, prior experiences and emotion regulation capacity (Weinberg & Mellano, 2023). This challenges simplistic interpretations of the *Yerkes–Dodson Law* and supports more ecological and multidimensional accounts of anxiety and performance (Hanoch & Vitouch, 2004). Consistent with this view, research examining test anxiety has identified a facilitative role for emotional arousal in the absence of intrusive cognitive worry, suggesting that anxiety can enhance performance when it remains within manageable bounds (Chin et al., 2017). When anxiety exceeds a learner's optimal zone, however, it is increasingly likely to trigger threat-based nervous system responses commonly characterised as fight, flight, freeze, or fawn (appease). These responses reflect automatic survival-oriented reactions to perceived danger rather than deliberate cognitive engagement. Fight responses may manifest as defensiveness, irritability, or over-verbalisation as individuals attempt to reassert control, while flight responses often involve avoidance behaviours such as withdrawal, disengagement, or a desire to escape the evaluative situation. As anxiety intensifies further, freeze responses may occur, characterised by mental blankness, reduced verbal output and impaired working memory, while fawn responses may involve excessive compliance or people-pleasing behaviours aimed at minimising perceived threat (Stafford, 2024).

In PBAs, which require spontaneous verbalisation, justification of reasoning and social interaction with an assessor, these threat responses are particularly consequential. When learners enter fight–flight–freeze–fawn states, cognitive processing shifts away from reflective and analytical thinking toward survival-oriented functioning, resulting in performance that may no longer accurately reflect underlying competence. Experimental evidence indicates that acute cognitive performance anxiety increases threat interference and impairs working memory, particularly among individuals with higher trait anxiety (Angelidis et al., 2019). This has direct implications for the validity and fairness of PBAs when anxiety is elevated beyond optimal levels. The point at which anxiety becomes counterproductive is especially variable among learners who experience social anxiety or who are neurodivergent. For these individuals, optimal arousal zones may be narrower and transitions into fight–flight or freeze responses may occur more rapidly in response to social, sensory, or evaluative demands. As a result, PBAs may escalate anxiety into threat-based states for some learners even when remaining facilitative or manageable for others, reinforcing the need to consider individual differences when evaluating assessment experiences. Accordingly, this study does not conceptualise anxiety as inherently harmful. Rather, it distinguishes between facilitative anxiety, which can support engagement and performance, and excessive or unhealthy anxiety that places learners in threat-based states and disrupts cognitive functioning. This distinction is particularly important in the context of PBAs, where elevated anxiety has been shown to interact with assessment design, learner characteristics and perceived stakes in complex ways (Heydarnejad et al., 2022). The focus of this research is therefore on anxiety that exceeds optimal arousal levels and undermines both learner experience and the validity of assessment outcomes.

The negative effects of heightened anxiety in PBAs

Especially for learners who experience social anxiety or who are neurodivergent, anxiety is recognised as one of the predominant unpleasant emotions encountered during PBAs (Burić & Frenzel, 2019; Shafiee Rad & Jafarpour, 2023). For those learners, anxiety is a condition characterised by fear, manifesting physiological responses such as an accelerated heart rate, hyperventilation, and perspiration (Eysenck, 1992). As previously discussed that while a manageable low-level of anxiety can sharpen focus and improve performance, for learners who experience social anxiety or who are neurodivergent, a high level of anxiety can overwhelm students and negatively impact outcomes in PBAs (Iqbal et al., 2010; Knight et al., 2013; Thamby et al., 2016). Although research shows that the use of PBAs are beneficial in reducing student anxiety when compared to traditional examinations (Heydarnejad et al., 2022), several studies show that the use of various assessment techniques within PBAs can still lead to an excessive high level of student anxiety, dissatisfaction and lack of motivation, for example in viva voce examinations (Hungerford et al., 2015; Iqbal et al., 2010; Scott & Unsworth, 2018). This is further reiterated by Chamorro-Premuzic et al. (2005) who suggest that within the various assessment techniques categorised as PBAs, oral and viva voce methods are the least preferred type of assessment for students. Furthermore, anxiety is identified as a primary cause for the lack of student motivation which is a condition crucial for enabling students to *"make certain academic decisions, participate in classroom activities, and persist in pursuing the demanding process of learning"* (Dörnyei, 2020, p. 2).

As student anxiety presents a key challenge in the effective delivery oral PBAs, there is a need to explore new approaches to delivering PBAs that reduce student anxiety but maintain the benefits of these assessment methods. Reducing student anxiety is not only desirable for wellbeing and inclusion but is also essential to preserve the validity and fairness of PBAs. Elevated anxiety has been shown to impair working memory, hinder retrieval and disrupt oral fluency (Aubrey, 2022; Bielak, 2025; Donate, 2022), which means that a students' performance may not accurately reflect their knowledge or competence. Moreover, excessive anxiety can reduce motivation and engagement (Cho et al., 2023; Heydarnejad et al., 2022), leading to long-term consequences beyond a single assessment event. Therefore, addressing anxiety is crucial to ensuring that PBAs remain both effective and equitable measures of learning.

Types of anxiety in PBAs

Research that focuses on test anxiety suggests it comprises multiple dimensions that interact to impair performance and emotional wellbeing. Cognitive anxiety includes intrusive thoughts and worries that interfere with concentration and working memory, often resulting in reduced task efficiency (Bielak, 2025; Scholing & Emmelkamp, 1990). Somatic anxiety refers to the physiological symptoms of arousal, such as increased heart rate, sweating, or trembling that frequently occur under test pressure and may exacerbate stress responses (Cho et al., 2023; Scholing & Emmelkamp, 1990). Social-evaluative anxiety, a distinct subtype, involves fear of negative evaluation by assessors or peers, particularly in face-to-face, high-stakes oral exams such as viva voce assessments (Kearney, 2013; Rao, 2020). In such scenarios, students often experience heightened stress due to direct scrutiny, which may lead to breakdowns in oral fluency and impaired retrieval (Aubrey, 2022; Donate, 2022). By contrast, the use of conversational agent for assessment or rehearsal may offer a less threatening interaction context, because it reduces real-time judgment and allows for controlled pacing and adaptive feedback.

AI conversational agents in education

The integration of AI conversational agents into educational processes presents a transformative opportunity to address pedagogical challenges, such as aligning AI-generated feedback with curricular expectations, ensuring reliability of assessment outcomes, and maintaining opportunities for authentic human dialogue. Emotional challenges also arise, including not only anxiety but also issues of trust, frustration when technical errors occur, and the possibility of over-reliance on mediated interactions, which may discourage students from developing coping strategies for the discomfort that authentic face-to-face assessments can create. By facilitating personalised and emotionally intelligent interactions during assessments, AI conversational agents can enhance learning and have the potential to significantly reduce student anxiety, fostering a supportive and engaging academic environment. Globally, educational frameworks are evolving to embrace a curriculum that is increasingly technology-enhanced and personalised. Artificial intelligence (AI) stands out as one of the key technological advancements of recent times, captivating the attention of educators, who particularly acknowledge the advanced capabilities of AI applications, such as generative AI, thereby seeing its adoption as a key enhancer of the efficiency and effectiveness of educational processes, such as assessments (Alshumaimeri & Alshememry, 2024). Among the various human–computer interaction (HCI) technologies used in education, AI conversational agents are emerging as a key area of growth and development. An AI conversational agent is a computer system that imitates natural conversation with human users through image, written, or spoken language (Laranjo et al., 2018). Since the first chatbot built in 1966, the evolution of computer science and AI research has resulted in AI conversational agents that can go beyond the facilitation of simple rule-based dialogue between computers and humans, to full AI conversational systems which have the ability to imitate human beings in aesthetic and voice, identify the intent of the user through minimal interactions and access various data sources to output conclusions, assumptions and decisions, sometimes upon which to take actions (Adamopoulou & Moussiades, 2020; Weizenbaum, 1966). Within the education sector, AI conversational agents are used in a myriad of contexts, including assessing learners', increasing peer dialogue, the provision of context for education such as role-play and answering student questions, among others (Gonda & Chu, 2019; Maryadi et al., 2017; Shorey et al., 2019; Wang et al., 2020). Any pedagogical context involving interactions between individuals or between humans and computers can be supported and/or enhanced using AI conversational agents, resulting in their popularity and acceptance. The conceptual framework developed as a result of a systematic review of pedagogical AI conversational agents by (Yusuf et al., 2025) suggest that contemporary agents can be used in a wide array of contexts within education, such as improving students' cognitive ability, providing them with pastoral care and giving instruction.

Recently, the emergence of generative AI through large language models (LLMs) have surged due to their advanced capabilities compared to earlier rule-based or narrow-domain conversational agents. An LLM is commonly defined as a mathematical model that predicts the likelihood of sequences of words or parts of words in a language, which serves as a foundation for computer software to carry out tasks that involve predicting text (Devlin et al., 2019). With the emergence of artificial neural networks, it has become feasible to create more sophisticated LLMs utilising transformer architecture, which further improves the ability of LLMs to understand and generate human-like text, including subtle emotional distinctions (Vaswani et al., 2017). This enables these models to better grasp the nuances of language, including context, tone, and the relationships between words, resulting in more accurate and coherent text processing and creation across a variety of applications such as language translation, question-answering systems, generative writing tools, and more recently multimedia development systems. An LLM-augmented AI conversational agent surpasses the operational boundaries of conventional agents, permitting users to initiate and sustain dialogues spanning an extensive array of subjects via open-dialogue, which significantly advances their applicability as comprehensive, multipurpose conversational systems (Abbasian et al., 2024; Cherakara et al., 2023). The use of an emotionally intelligent LLM-augmented pedagogical AI conversational agent has the ability to enhance assessments and potentially improve student anxiety.

AI conversational agents to reduce student anxiety

Before turning to education, it is useful to note that AI conversational agents have already demonstrated strong potential in other anxiety-inducing domains such as healthcare and business. These contexts, like education, require trust, emotional resilience and high-stakes decision-making, hence making them informative analogues. AI conversational agents have proven effective in improving the psychological well-being of users in various fields by providing personalised, evidence-based support. Empathy-driven tools such as Wysa and Woebot alleviate symptoms of depression and loneliness through judgement-free interactions and cognitive behavioural therapy (Fitzpatrick et al., 2017; Inkster et al., 2018; Loveys et al., 2019). Multimodal communication enhances user engagement, creating more human-like connections that foster emotional resilience (Abd-Alrazaq et al., 2019). Studies also show that agents are able to improve self-care behaviours (Tawfik et al., 2023). The use of agents in these non-educational domains have transferable mechanisms which can be applied to educational contexts, such as empathetic interaction and stress management that can be applied to academic scenarios where student anxiety is prevalent.

Specifically, within the context of education, AI conversational agents have shown significant promise in reducing student anxiety and enhancing emotional well-being. Research demonstrates their ability to address both psychological and academic challenges effectively. Klos et al. (2021) found that using an AI conversational agent significantly reduced anxiety symptoms in university students, highlighting the effectiveness of technology-based engagement compared to traditional resources. Similarly, Liu et al. (2022) reported that chatbot-based interventions outperformed bibliotherapy in reducing depression and anxiety, particularly during initial phases, while fostering strong therapeutic alliances. Wang et al. (2024) emphasised the dual role of AI conversational agents in improving learning outcomes and supporting well-being. Their ability to provide empathetic and context-specific interactions underpins their holistic impact on student experiences. This is further evidenced by El Shazly (2021), who found AI conversational agents beneficial in managing foreign language anxiety. Hawanti and Zubaydulloevna (2023) observed that LLM-augmented AI conversational agents reduced anxiety in English writing classes by offering instant feedback and fostering a supportive, paced learning environment. These findings underline the value of AI conversational agents as accessible and personalised tools for addressing anxiety and promoting student success.

Limited research has been conducted in the context of using AI conversational agents to reduce student anxiety in PBAs. The study by Alcorn and Cheesman (2022) introduces a technology-assisted viva voce exam (TaVIVA) aimed at reducing student anxiety during oral PBAs in pharmacy education. By utilising an AI conversational agent to deliver questions in various multimedia formats, such as text, audio, and audiovisual, students engage in a controlled environment without the immediate presence of assessors which typically heightens anxiety. This approach allows students to respond in a comfortable setting, record their answers, and review them later, leading to a more positive assessment experience. While the findings indicate reduced anxiety and positive student feedback regarding fairness and functionality, the study presents opportunities for further exploration. These include providing more detailed information about the technology-assisted platform used and assessing its impact on usability. Furthermore, according to the study, addressing potential technical challenges associated with the technology-assisted viva voce exam would present an opportunity to refine the assessment experience and ensure its reliability in diverse educational contexts.

The potential of emotionally intelligent, LLM-augmented AI conversational agents to reduce student anxiety in PBAs

PBAs represent a significant difference over traditional examination methods in pedagogy (Gallardo, 2020). However, they introduce challenges to student wellbeing, notably heightened levels of anxiety, leading to lack of motivation, especially for learners who experience social anxiety or who are neurodivergent (Burić & Frenzel, 2019; Shafiee Rad & Jafarpour, 2023). In this context, AI conversational agents present a promising avenue for facilitating PBAs. When these agents are augmented with LLM capabilities, they have the potential to evolve into emotionally intelligent systems. Such systems can recognise and respond to the emotional state of students, thereby potentially reducing student anxiety during PBAs. This area holds considerable promise and warrants further investigation. There has been some research on the use of AI conversational agents in the context of assessments (Gonda et al., 2018; Lee & Fu, 2019; Maryadi et al., 2017), and further research conducted in the context of an AI conversational agents being used to improve student well-being, as outlined above. However, research exploring emotionally intelligent LLM-augmented AI conversational agents specifically designed for pedagogical purposes, and improving the student experience during assessment, remains limited. There is a need for new research that explores the potential of emotionally intelligent LLM-augmented pedagogical AI conversational agents, used within PBAs, to fully understand their benefits and implement effective technology-assisted assessment solutions. Such agents have the potential to play a pivotal role in mitigating student anxiety within the PBA setting.

Theoretical underpinning for the reduction of anxiety when adopting technology

A logic model that structures how technology may reduce anxiety in viva voce assessments is presented in Fig. 1.

In this model, *Inputs* refer to the contextual and design factors that shape the assessment environment. *Theoretical Mechanisms* represent the established psychological and educational theories that explain how these inputs influence emotional and cognitive processes. *Mediators* are the intermediate processes or states, for example, reductions in perceived social threat or improvements in self-regulation, that link the theoretical mechanisms to the final outcomes. *Outcomes* are the intended effects on learners, such as lower anxiety, increased cognitive availability and more effective demonstration of competence. The model begins with a set of *inputs*, including the role of *technology mediation*, the *assessment context* and *learner factors*. These inputs interact to shape the learners' experience of

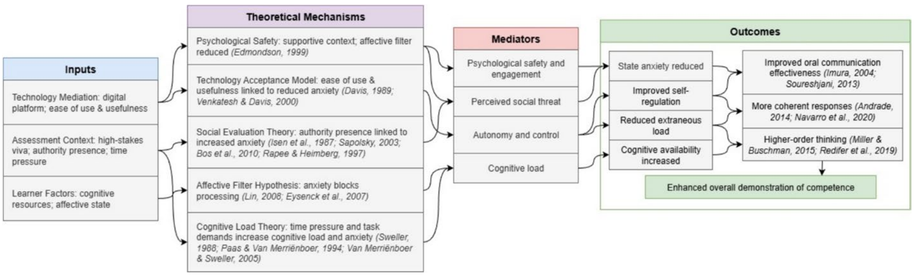

Fig. 1 A Logic Model of How Technology May Reduce Anxiety in PBAs

performance-based assessments. The influence of these inputs can be explained by several *theoretical mechanisms*.

From a psychological perspective, *social evaluation theory* suggests that the presence of an evaluative authority can heighten self-consciousness and apprehension, leading to elevated anxiety (Isen et al., 1987; Sapolsky, 2003). This is especially salient in viva voce formats, where responses must be articulated spontaneously and under strict time constraints, which increases both *cognitive load* (Sweller, 1988) and affective interference (Eysenck et al., 2007). Consistent with this, the *affective filter hypothesis* (Lin, 2008) adds that heightened anxiety can block processing and retrieval, limiting the demonstration of competence. By contrast, the *Technology Acceptance Model* (Davis, 1989) highlights how perceptions of ease of use and usefulness in technology-mediated environments can reduce anxiety and foster acceptance (Venkatesh & Davis, 2000). Finally, theories of *psychological safety* emphasise the importance of a supportive, consistent, and predictable interactional context, which can lower the affective filter and promote engagement (Edmondson, 1999; Pekrun, 2006).

These theoretical mechanisms translate into *mediators* that alter the dynamics of the viva experience. Technology-mediated assessment environments have the potential of reducing direct exposure to authority figures, thereby lowering *perceived social threat* and decreasing state anxiety (Bos et al., 2010; Rapee & Heimberg, 1997). Allowing students to pace, pause or reflect before responding increases *autonomy and control*, thereby creating a more favourable environment for learning and performance (Jiang et al., 2019; Reeve, 2009). Such features have also been associated with reduced *cognitive load* and improved self-regulation (Kalyuga, 2008; Paas & Van Merriënboer, 1994; Van Merrienboer & Sweller, 2005). Together, these mediators foster increased *psychological safety and engagement* while freeing up cognitive resources that would otherwise be consumed by managing anxiety. The result is a set of *outcomes* that fundamentally reshape the learners' experience of assessment. These include *state anxiety reduced*, which allows students to approach questions without the paralysing effects of fear or worry; *improved self-regulation*, enabling them to monitor and control their emotional responses in real time; *reduced extraneous load*, meaning that fewer cognitive resources are wasted on stress management or irrelevant distractions; and *cognitive availability increased*, freeing working memory capacity to focus on the demands of the task itself. Together, these outcomes create the psychological and cognitive conditions necessary for students to engage more productively and perform more effectively during oral assessments.

Building on these immediate changes, the model also identifies a second set of outcomes that connect emotional regulation and cognitive resources to observable performance behaviours. When anxiety is reduced and self-regulation is improved, students are more likely to demonstrate *improved oral communication effectiveness*, articulating their ideas clearly and with confidence (Imura, 2004; Soureshjani, 2013). Further, when students are freed from intrusive thoughts or excessive pressure, they are also able to provide *more coherent responses*, structuring arguments logically and responding directly to examiners' prompts (Andrade, 2014; Navarro et al., 2020). In addition, the availability of cognitive resources facilitates *higher-order thinking*, allowing students to move beyond surface-level recall and towards creative thinking, critical analysis, evaluation and synthesis (Miller & Buschman, 2015; Redifer et al., 2019). Collectively, these outcomes culminate in an *enhanced overall demonstration of competence*, which is the central goal of viva voce assessments. In effect, by creating conditions in which learners can fully access and communicate their knowledge and skills, this model suggests that technology could support students' anxiety management (Danieli et al., 2022; De la Puente et al., 2024; Li & Jan, 2023; Manole et al., 2024), allowing them to be redirected towards higher-order thinking, coherent structuring of responses and more effective oral communication.

Aims of this study

Prior studies have been conducted which indicate that AI conversational agents have the potential of improving students' performance (Lee et al., 2022; Xu et al., 2024). Therefore, this study presents "AIvaluate", an emotionally intelligent, LLM-augmented pedagogical AI conversational agent developed with the aim of facilitating oral PBAs for students, towards reducing student anxiety and improving student experience. The AIvaluate system consists of innovative functionalities designed to address student anxiety in PBAs. Specifically, the following research questions are addressed in this study:

> RQ1. To what degree does AIvaluate reduce students' anxiety levels during PBAs compared with equivalent traditional face-to-face assessments?
> RQ2. How satisfied, in terms of usability, are students with the AIvaluate assessment compared with equivalent the traditional face-to-face assessment format?
> RQ3. What factors influence students' preferences for AIvaluate assessments compared with traditional face-to-face assessments?

The remainder of this paper presents the AIvaluate system architecture and application walkthrough in Sect. "The AIvaluate system", the methods used for empirical evaluation of the system is presented in Sect. "Methods", followed by the results based on the data collected in Sect. "Results". Furthermore, a discussion of the results and future recommendations is presented in Sect. "Discussion" and conclusions are drawn in Sect. "conclusion".

The AIvaluate system

This section presents an overview of AIvaluate, including the system architecture, workflow, and operational mechanics. It further provides a step-by-step walkthrough of the application from the perspective of student users, demonstrating how they interact with the system during AI conversational agent-based assessments.

System architecture and workflow

AIvaluate is an emotionally intelligent LLM-augmented pedagogical AI conversational agent, deployed as a disembodied chatbot which has three key functionalities; 1) a *conversational interface* wherein dialogue can take place between a student and teacher during oral PBAs, 2) *emotional state detection* for the student to report levels of anxiety and 3) *LLM-augmentation* which uses the students' emotional state data to generate automated suggested replies for the educator, who can then provide contextually relevant and emotionally intelligent responses to reduce student anxiety levels. A high-level system architecture and data flow model for AIvaluate is presented in Fig. 2.

The *conversational interface* is developed to be accessible within a browser environment. The first feature that the AIvaluate system hosts is an interface wherein students can converse with teachers through inputs and outputs. The input for students can be text-based or verbal, in the latter case, the application utilises a dictation *speech-to-text (STT)* library to ensure the input at terminal level is string datatype. Outputs for students are also displayed in textual format within the interface, in addition to utilising a *text-to-speech (TTS)* library to convert the text into audio (spoken form). Utilisation of these technologies presents the student with an anthropomorphic conversational experience. For teachers, the inputs are LLM-generated suggested adaptive responses, and similar to students, the outputs are also in text and spoken form. The *student emotional state detection* feature of AIvaluate is a *self-reporting tool* which is used by the students. The *self-reporting tool* input is presented as a slider to the student which, during an active session, allows them to input their emotional state from a scale-slider between 1 and 10, indicating their level of anxiety. Additionally, a session *log* is produced for prospective analysis. This real-time emotional state data is utilised by the *LLM-augmentation* integration using a custom JSON-based *Application Programming Interface (API)*. The *self-reporting tool* was developed using object-oriented technologies for server-side operations with a relational *database* as the storage medium for logging data, and front-end technologies for browser-side behaviour manipulation.

Another feature of AIvaluate is the *LLM-augmentation*, which is used in the *teacher-interface* to generate emotionally intelligent and automated *suggested adaptive responses* for replying to student inputs. The LLM utilised is OpenAI's Generative Pre-trained Transformer (GPT-4) which is incorporated into the agent using their API integration library, which allows requests to take advantage of OpenAI's trained neural network to produce

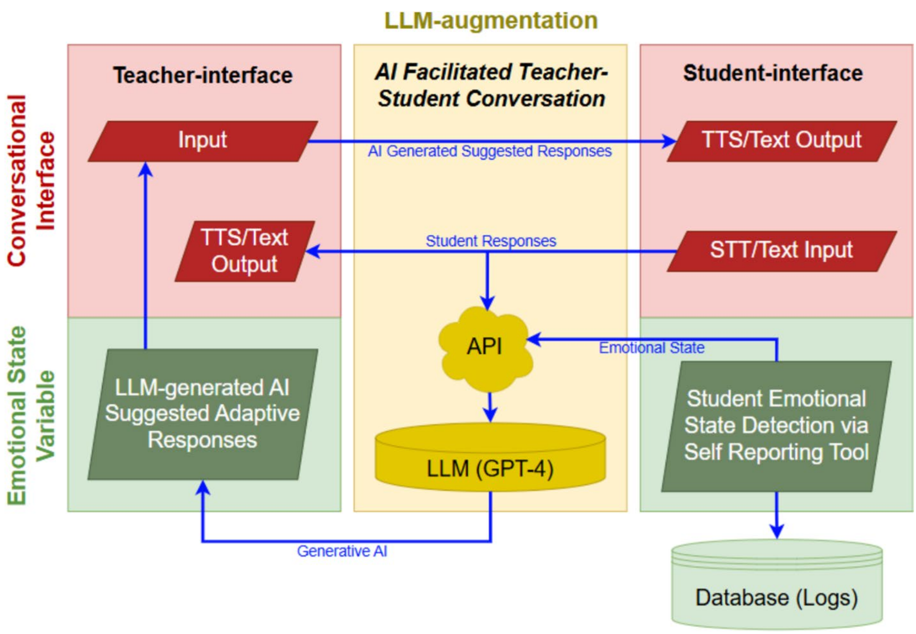

Fig. 2 AIvaluate system architecture and workflow model

relevant, coherent, and contextually appropriate, adaptive responses. In addition to the LLM producing a suggested response for the teacher based solely on the students' inputs, the *LLM-augmentation* codebase includes a custom variable to report the students' real-time emotional state based on the live feed of the *self-reporting tool*. The emotional state data is then used by the LLM to generate emotionally intelligent responses to the student.

System walkthrough

Upon launching AIvaluate within a full-screen browser environment, the student is welcomed with an introduction screen, which is presented in Fig. 3.

The introduction *launch screen* welcomes the user into the system and allows the user to select their role using the *user type selection* menu, namely a teacher or student. Upon selecting the *student* user type in the introduction screen, the student is transitioned to the *student interface* screen, which is presented in Fig. 4.

The *student interface* screen hosts a *chat history panel* where the student can observe the inputs that have already been received from the teacher. The chat panel employs TTS functionality to convert the educator's messages into audio-based verbal prompts. Positioned beneath this is the *student input selection area*, which enables the student to compose and dispatch messages to the teacher. This input area is equipped with *controller buttons*, including options to send a completed message or clear the text field. The *dictation functionality* allows the student to utilise the STT feature via microphone which allows verbal input of responses, thereby augmenting the naturalistic communication experience. Research highlights that modern TTS/STT systems are perceived as natural and intelligible, fostering a more anthropomorphic and engaging environment (Skantze & Hjalmarsson, 2010). By providing information through both visual and auditory channels, TTS/STT reduces cognitive strain, supports inclusivity and contributes to a psychologically safer assessment experi-

Fig. 3 AIvaluate introduction screen

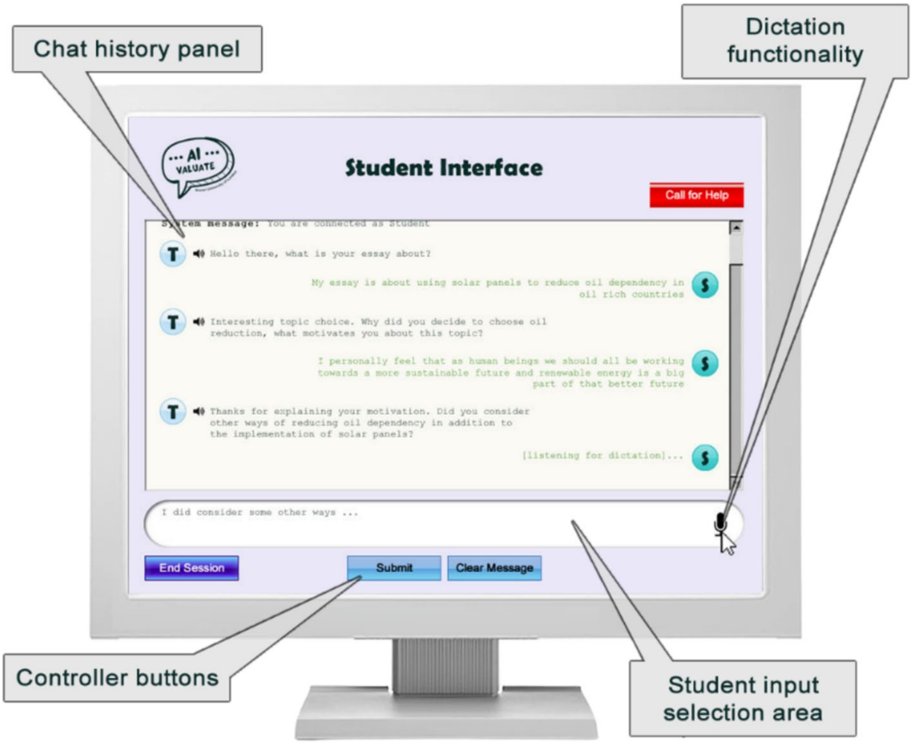

Fig. 4 AIvaluate student interface screen

ence, key factors in lowering student anxiety during oral PBAs. In addition to this interface, the student also has access to the emotional state reporting tool, which is available to the student as a secondary display on a touch-screen tablet. The emotional state reporting tool is presented in Fig. 5.

The *emotional state reporting tool* hosts the student anxiety level *self-reporting input*, a critical component that allows the student to actively monitor and report their live emotional state throughout the assessment session. The input features a *slider bar*, enabling the student to self-assess and indicate their current level of anxiety on a scale from 1 to 10, with descriptors and semiotic *emoticons* to support their understanding of the levels. This input operates in real-time, allowing the student to make continuous adjustments to reflect their emotional state at any given moment which is displayed in the *selection view*, thereby providing a dynamic and immediate self-reflective assessment that informs the ongoing interaction. The slider was chosen because continuous self-report tools are more sensitive than categorical scales, capturing subtle, moment-to-moment fluctuations in anxiety. They are intuitive to use, impose minimal cognitive load and generate fine-grained real-time data that supports adaptive system responses (Krieglstein et al., 2022). From the introduction screen as shown in Fig. 3, if the user selects the *teacher* user type, the teacher will be transitioned to the teacher interface screen, which is presented in Fig. 6.

The *teacher interface* hosts a *chat history panel* where the teacher can observe the inputs received from the student. Like the student interface, the chat panel employs TTS function-

Fig. 5 AIvaluate emotional state reporting tool

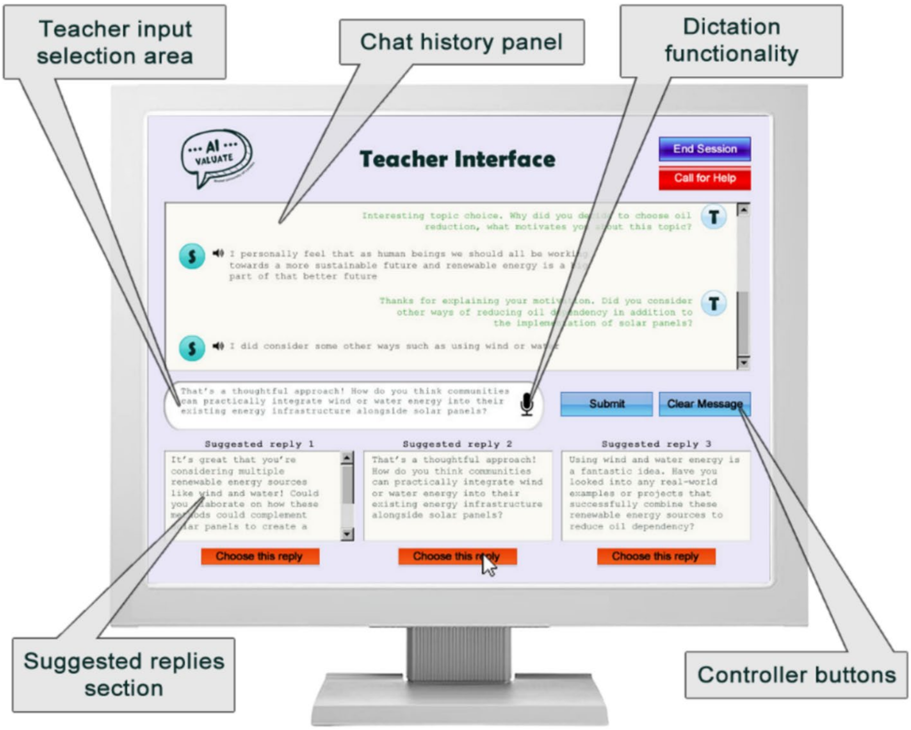

Fig. 6 AIvaluate teacher interface screen

ality to convert the student's messages into audio-based verbal prompts. Positioned beneath this is the *teacher input selection area*, which enables the teacher to compose and dispatch messages to the student. This input area is equipped with *controller buttons*, including options to send a completed message or clear the text field. The *dictation functionality* allows the teacher to utilise the STT feature via microphone which allows verbal input of responses. In addition to this interface, the teacher also has access to a *suggested replies section*, wherein the system generates three emotionally intelligent, suggested adaptive responses for the teacher, based on three variables; the student's last input, the conversation history for context and the student's emotional state based on the self-reporting tool. Teacher-suggested responses can positively influence student experience by reducing cognitive load and supporting students' confidence during challenging tasks. Research shows that when learners receive guided prompts, they report greater clarity, engagement and reduced anxiety, which contributes to a more supportive and motivating learning environment (Er et al., 2025).

Methods

This study employed a quasi-experimental, within-subjects design to evaluate student experiences across two assessment conditions: a traditional face-to-face viva and an AI-led viva using AIvaluate2. Each participant experienced both formats in a counterbalanced order, enabling within-subject comparison and control for order effects. Data was collected from three complementary sources: (1) log data of students' self-reported emotional states during assessments, (2) post-assessment usability questionnaires, and (3) open-response reflections. While the primary design was quasi-experimental, incorporating multiple data types enabled a broader understanding of both the measurable and subjective aspects of the student experience. Figure 7 provides an overview of the protocol.

Participants

A total of 35 students were recruited from a British international school. These students were in their final year of the International Baccalaureate Diploma Programme (IBDP), specifically engaged in completing their Extended Essay (EE), which includes a summative PBA known as the *viva voce*. The number of student-participants required for this study was estimated using G*Power 3.1 software. To ensure a power of 0.80 with a medium effect size of 0.5 (dz), the calculation determined that a minimum total of $N=34$ participants would be required. The inclusion criteria for student participants were 1) age (17–18 years), 2) level of study (pre-university programme), 3) completion of the EE and pending the viva voce, 4) computer literacy, and 5) native-level proficiency in the English language. All students met the required criteria. Additionally, 35 qualified teachers were recruited from the same institution to conduct the assessments with a student–teacher ratio of 1–1. No participants reported any functional impairments that would affect their ability to engage with the AI agent, although minor adjustments to software settings were made to accommodate user preferences (such as adjusting audio output levels and screen size/scale levels).

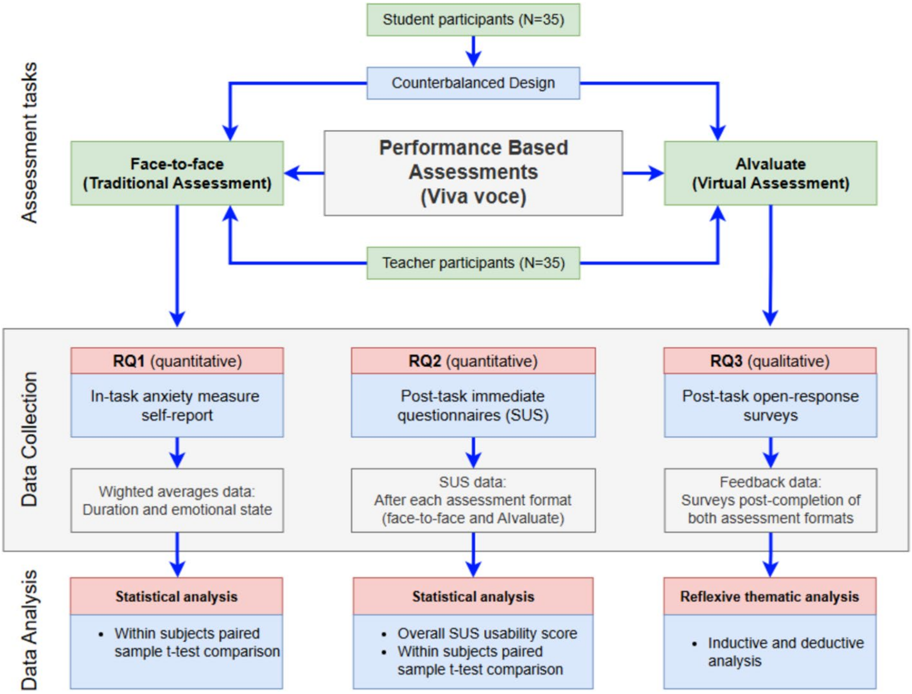

Fig. 7 Data collection and analysis procedure

Assessment objectives and performance measures

The viva voce assessment in this study was aligned with the IBDP Extended Essay (EE) assessment framework, which requires students to demonstrate a synthesis of subject knowledge, critical analysis and communication skills. Specifically, the assessment aimed to evaluate the following objectives:

1. **Conceptual understanding and subject knowledge**; the extent to which the student could explain the research topic, justify methodological choices and articulate key findings in relation to the EE subject area.
2. **Critical thinking and evaluation**; the ability to reflect on strengths and limitations of the research, discuss implications and identify areas for further exploration.
3. **Organisation and coherence of responses**; the logical structuring of explanations, clarity of argumentation and the ability to respond to examiner prompts in a coherent manner.
4. **Oral communication skills**; clarity of speech, linguistic precision and appropriate use of subject-specific terminology.

These objectives were included to provide context for interpreting the emotional state data. While examiners naturally considered these domains during the viva voce, the present study did not record or analyse formal performance scores, as the primary focus was on measuring emotional and usability outcomes. Future research could explore the interaction between specific performance dimensions and emotional responses.

Protocol and instrumentation

Interactive PBA-based viva voce assessment sessions were held with all participants as part of the study on the effectiveness of the AIvaluate system. Each student participated in two distinct assessment sessions, one face-to-face and one virtual. A researcher was present during all sessions, and written consent was obtained from all participants and guardians as required. On arrival at each session, any questions from participants were addressed, and a brief presentation was provided detailing the research process. Each student participated in two assessment meetings with the same teacher, separated into face to face and virtual formats, and a counterbalanced design was employed (i.e. alternating order of meetings for student participants) to mitigate order and carryover effects and increase the generalisability of the research (Brooks, 2012).

Face-to-face assessment meeting

The face-to-face assessment involved a conventional viva voce session conducted in person, lasting 10–15 min, where the student and teacher engaged in a discussion about the student's EE. During this session, the students were required to explain, justify, and reflect upon various aspects of their essay, responding to questions posed by the teacher. To maintain consistency and standardisation of interactions, the teacher was provided with a simple script to guide the initial stages of the conversation and ensure comparability across sessions. Throughout the face-to-face assessment, students were asked to self-report their anxiety levels using an emotional state self-reporting tool, which is the same tool used for the virtual assessment as presented in Fig. 4, designed to capture real-time emotional states during the interaction.

Virtual assessment meeting

The virtual assessment with students was conducted using the AIvaluate system, designed to simulate an AI-augmented interaction through the Wizard of Oz technique. In this setup, students were unaware that the teacher was the interlocutor; instead, the assessment was framed as being conducted by AIvaluate, a fully automated artificially intelligent conversational agent (Steinfeld et al., 2009). The sessions were hosted in separate, soundproof rooms to prevent students and teachers from hearing each other during the interaction. Figure 8 illustrates the room layout and hardware setup, showing how the student was situated, and the equipment arranged.

The *teacher room* hosts the teacher who was an interlocutor and conversed through AIvaluate to facilitate the assessment. In the *Student room*, the student interacted with AIvaluate using a setup which consisted of two windows; one labelled *Student View* displaying the AIvaluate interface and another enabling the student to input their emotional state (labelled *ES-I*). The emotional state reporting input sent data using the *API* as detailed in Sect. "System architecture and workflow". The AIvaluate system was designed to track and log emotional states in real time, providing critical data for analysing anxiety levels during the virtual assessment. The virtual assessment, similar to the face-to-face session, lasted 10–15 min.

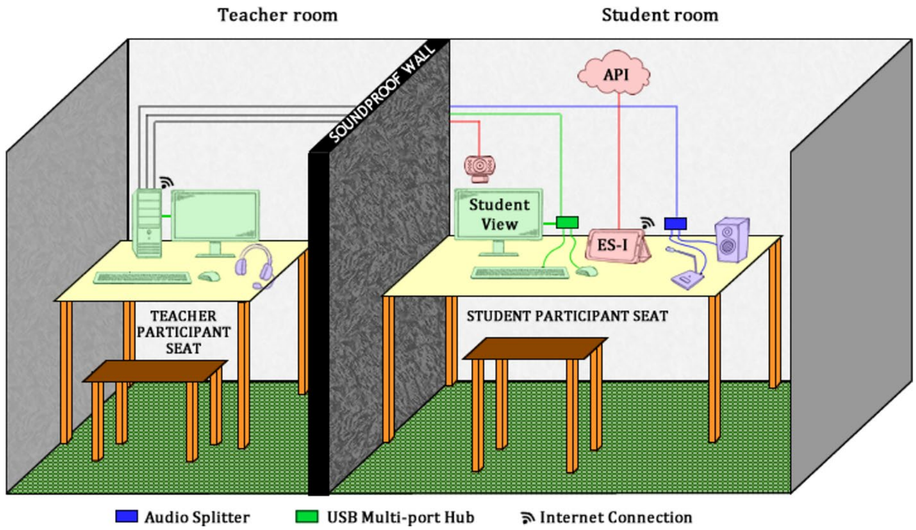

Fig. 8 AIvaluate Hardware Setup

Instruments for data collection

A combination of digital and survey-based instruments was used to capture students' emotional states, perceptions of usability and subjective experiences across both assessment conditions.

Emotional state self-reporting tool: To measure real-time fluctuations in anxiety, students used a custom-built digital self-report slider integrated into the AIvaluate interface. This instrument enabled participants to continuously indicate their anxiety level on a 10-point scale (1 = extreme anxiety, 10 = no anxiety/very calm). Each input was timestamped and logged, generating high-resolution data that captured the trajectory of emotional states throughout the assessment session.

System usability scale (SUS): Following each assessment format, students completed the SUS questionnaire, a widely recognised and validated 10 item tool used for assessing subjective usability and user satisfaction (Brooke, 2013). In addition to the 10 standard SUS items, four additional bespoke statements relating to evaluating the student experience were added. Each statement was rated on a 5-point Likert type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Each SUS item was modified in accordance with the practitioner guidelines to reflect the specific context of the assessment by replacing the word "system" with either "face-to-face assessment" or "AIvaluate virtual assessment" depending on the participants assessment type. This adaptation ensured that the scale accurately measured the participants' subjective satisfaction and acceptance of the assessment format.

Qualitative open-response questions: In addition to quantitative instruments, a short survey comprising open-response prompts was administered after both assessment conditions. These questions sought to capture richer reflections on students' experiences, focusing on usability, acceptability, and the perceived emotional impact of each format.

Together, these instruments allowed for triangulated measurement of affective, usability and experiential dimensions, providing complementary perspectives on the comparative effectiveness of the two assessment formats in relation to student experience and anxiety.

Data collection procedure

Data was collected across three sequential phases to ensure a comprehensive understanding of students' experiences in both assessment conditions.

Phase 1, Emotional state logs: During each assessment, students reported their moment-to-moment anxiety levels using the digital self-report slider. These data were automatically timestamped and logged within the system, providing a continuous record of emotional fluctuations throughout both the AIvaluate and face-to-face sessions.

Phase 2, Post-assessment questionnaires: Immediately following each assessment, students completed the SUS, along with the additional bespoke items targeting specific aspects of student experience. This provided quantitative measures of usability and acceptability for both formats.

Phase 3, Open-response reflections: Finally, participants responded to open-ended survey questions after they had completed both assessment formats, which were designed to capture deeper reflections on their experiences. These responses yielded qualitative insights into perceived fairness, comfort and emotional impact across the two assessment contexts.

By structuring the collection into three phases; real-time logging, post-task usability ratings, and reflective commentary, the study captured both immediate and reflective perspectives, allowing for triangulated analysis of student anxiety, usability, and experience.

Data analysis

This section is ordered according to the three primary data collection phases, present in Sect. "Instruments for data collection", which includes analysis of the students' emotional state input data, SUS questionnaire data, and qualitative survey response data.

Students emotional state input data analysis (RQ1)

As discussed in Sect. "Protocol and instrumentation", the emotional state reporting feature of AIvaluate allows the student to use a scale-slider between levels 1 and 10, indicating their level of anxiety at any given time during the assessment. The system logs the timestamp and the level of anxiety into the database, which is used for analysis. For each student, during both the face-to-face and virtual assessment sessions, a weighted mean was calculated which considered the level of anxiety reported by the student (emotion code), and the duration of each emotional state. To calculate the weighted mean, the emotion code was multiplied by the duration (in seconds) that the participant remained in that state. These weighted values were then summed across all recorded emotions for a given session. The total duration across all emotional states was also summed. Finally, the sum of the weighted emotion codes was divided by the total duration sum, producing a weighted average emotional state that accurately reflected the student's overall anxiety level during each assessment condition. This method allowed for a more precise and representative measure of anxiety fluctuations over the course of each assessment. The mean and standard deviation of all the sessions in both the face-to-face and virtual assessments were calculated to provide a comparative benchmark between the two assessment types. Additionally, the paired-sample t-test was used to compare the weighted average emotional state values for both the virtual and face-to-face assessments, enabling an assessment of anxiety level differences and determining the statistical significance of any observed variation between the two conditions.

SUS questionnaire data analysis (RQ2)

Overall SUS scores were calculated and interpreted according to the SUS acceptability range, and the adjective and school grading scales (Brooke, 1996). This involved calculating a mean SUS representative value on a 100-point rating scale for each sample. These scores were then mapped to descriptive adjectives (Best imaginable, Excellent, Good, OK, Poor, Worst Imaginable), an acceptability range (Acceptable, Marginal-High, Marginal-Low, Not acceptable) and a school grading scale (i.e. 90–100 = A, 80–89 = B etc.). The baseline adjective and acceptability ranges are derived from a sample of over 3000 software applications (Bangor et al., 2009). To further analyse the usability of each assessment format, paired-sample t-tests were conducted to determine whether there were statistically significant differences between the mean SUS scores for face-to-face and virtual assessments on each individual SUS item. IBM SPSS statistics package version 29.0.1 was used for this analysis. This approach allowed for a detailed item-by-item comparison, examining how each usability aspect, as represented by the SUS items, differed between the two assessment types. By using paired-sample t-tests, we accounted for the fact that the same participants provided SUS ratings for both formats, thus controlling for individual variability and increasing the precision of the comparisons. Each t-test assessed whether the mean difference in scores for a specific SUS item, such as ease of use, consistency, and confidence using the assessment, was statistically significant, thereby offering insights into specific usability strengths or weaknesses unique to each format.

Qualitative open-response survey data analysis (RQ3)

Reflexive thematic analysis, using NVivo software package version 12.6.1.97, was conducted to analyse the open-response survey data (Braun & Clarke, 2006, 2012, 2013, 2019, 2021). Both an inductive and deductive approach to analysis was adopted, as coding and analysis often do not fit neatly into a single approach; instead, they typically involve a blend of both methods (Byrne, 2022). The integration of top-down (deductive) and bottom-up (inductive) coding processes in this approach facilitates a comprehensive and rigorous analysis (Proudfoot, 2023). Themes in deductive analysis were a priori researcher-led, while the inductive analysis allowed for new themes to emerge from the data (D'Amore et al., 2023; Tafazoli & Meihami, 2023). This process involved iterative and collaborative coding of survey responses, allowing themes and sub-themes to emerge reflexively across repeated analysis cycles, providing valuable contextual insights into the emotional and usability-related dimensions of both assessment types, which allowed for an in-depth exploration of students' subjective experiences and perceptions of each assessment condition. While inter-coder reliability was not calculated, this is consistent with the principles of reflexive thematic analysis, which conceptualises coding as a subjective, interpretative, and contextually situated process rather than a reliability test (Braun & Clarke, 2021).

Two researchers independently engaged with the data, generating initial codes and noting patterns of meaning. Through iterative cycles of discussion, disagreement, and refinement, the team collaboratively developed a robust thematic framework. This reflexive dialogue enhanced the depth and transparency of interpretation. Data saturation was assessed by tracking the emergence of new codes and sub-themes across responses. No new substantive categories emerged in the final set of responses, and thematic coverage was observed across participants, indicating sufficient saturation for the scope of this analysis. Memo writing and cross-case comparisons supported thematic consistency. This approach allowed an in-depth exploration of students' subjective experiences and perceptions of each assessment condition to be carried out. To illustrate the analytic coding process, Table 1 presents an example of how raw data extracts are coded and then clustered into a sub-theme and theme. This demonstrates the movement from initial codes to higher-level categories, a process consistent with reflexive thematic analysis. This example is illustrative rather than exhaustive, intended to demonstrate the coding process.

The quotes selected to represent a sub-theme are chosen because they collectively capture the essence of that sub-theme and contribute to the overarching theme. Other similar quotes are not included as they reiterate patterns already exemplified by the selected extracts.

Table 1 Example of the thematic coding process

Raw quote extracts	Code(s)	Sub-theme	Theme
"I felt like there was a lot more pressure on me in the face-to-face meeting" (Participant 8)	Increased pressure in F2F	Reduced Pressure in AI-led format	Strengths of technology-led assessments
"I didn't have to feel the pressure of explaining in front of the teacher" (Participant 35)	Less visible evaluative cues	Reduced Pressure in AI-led format	Strengths of technology-led assessments
"There was less judgement because, frankly, I couldn't see the teacher's reactions to my response" (Participant 16)	Absence of judgment from teacher	Reduced Pressure in AI-led format	Strengths of technology-led assessments
"I didn't feel like I was being judged for every word I said, so I could explain myself more calmly" (Participant 32)	Increased calmness in expression	Reduced Pressure in AI-led format	Strengths of technology-led assessments
"It felt less intense talking to AI because I wasn't worried about the teacher's reactions" (Participant 6)	Reduced intensity	Reduced Pressure in AI-led format	Strengths of technology-led assessments

Table 2 Student anxiety levels for aivaluate and face-to-face assessments

Face to face		AIvaluate		Gap score	df	t	p (1-tail)
M	SD	M	SD				
5.86 [a]	1.47	6.50 [a]	1.24	0.64	34	–1.97	.028*

Format	Min statistic	Max statistic	Skewness		Kurtosis	
			Statistic	SE	Statistic	SE
Face-to-face	3.23 [a]	9.00 [a]	.28	.40	–.40	.79
Aivaluate	3.97 [a]	8.86 [a]	–.16	.40	–.54	.79

[a] Anxiety Levels on a scale of 1–10, with 1 = "Very Anxious" and 10 = "Very Calm".

* Statistically significant < 0.05 level.

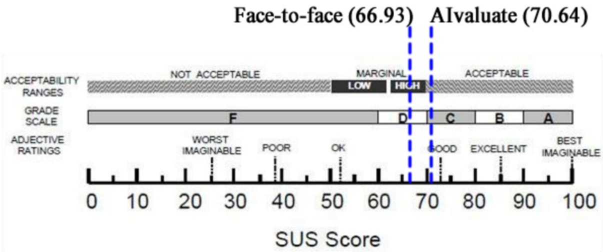

Fig. 9 SUS rating scale results for aivaluate and the face-to-face assessment

Results

Student anxiety

The first research question is to compare PBAs conducted through AIvaluate with traditional face-to-face assessments to identify the degree of reduction in students' anxiety levels, to identify whether there are any significant differences. The results of the comparison of the anxiety levels of students between AIvaluate and the face-to-face assessments are presented in Table 2.

The result of the paired samples t-test comparing student anxiety levels between AIvaluate and face-to-face assessments reveals that students reported significantly lower calmness levels during the AIvaluate assessments (M = 6.50, SD = 1.24) compared to the face-to-face assessments (M = 5.86, SD = 1.47), $t(34)$ = -1.97, p = .028 (*1-tailed*). The gap score (M = 0.64) highlights this difference. Skewness and kurtosis statistics suggest relatively normal distributions for both formats, with AIvaluate showing a slightly negative skew (-0.16) compared to face-to-face assessments (0.28). These findings indicate a statistically significant difference in anxiety levels, with AIvaluate hosted virtual assessments being associated with reduced anxiety.

System usability

The second research question aims to evaluate and compare the usability and user experience of AIvaluate and face-to-face assessment formats using the System Usability Scale (SUS). The overall SUS scores are illustrated in Fig. 9, which shows that the score for AIvaluate is 70.64, which, according to the SUS evaluation framework, corresponds to "good" (descriptive adjective), "acceptable" (acceptability range), and a "grade C" (school grading scale). In comparison, the face-to-face assessment scored slightly lower than AIvaluate, with an overall SUS score of 66.93, also rated as "good", "marginal" and a "grade D". Findings suggest that AIvaluate is more acceptable in terms of usability in comparison to the face-to-face assessment.

Follow-up analysis of individual SUS items is conducted to evaluate the usability of face-to-face assessments and AIvaluate assessments. Paired-sample t-tests are used to compare the mean SUS scores for each format across individual SUS items. Table 3 presents the individual SUS item results, differences (denoted as Gap Score) and corresponding significance values.

The negative SUS items (S2, S4, S6, S8, and S10) are reversed so that scores indicate a positive response to allow comparability. SUS items S1-10, analysis reveals statistically

Table 3 AIvaluate and face-to-face assessments comparison of SUS scores

SUS items	Face-to-face M	AIvaluate M	Gap Score	df	t	p (1-tail)
S1: I think I would like to have face-to-face/Aivaluate assessments frequently	3.51	3.17	-0.34	34	1.15	.129
S2: I found the face-to-face/Aivaluate assessment unnecessarily complex. [a]	3.86	3.94	0.09	34	–0.38	.353
S3: I thought the face-to-face/Aivaluate assessment was easy to participate in/use	4.06	4.03	-0.03	34	0.12	.452
S4: I think that I would need the support of a teacher/technical person to be able to understand/use the face-to-face/Aivaluate assessment.[a]	3.54	3.97	0.43	34	–1.74	.046*
S5: I found the various aspects/functions of the face-to-face/Aivaluate assessment were well integrated	3.80	3.60	-0.20	34	0.94	.176
S6: I thought there was too much inconsistency in the face-to-face/Aivaluate assessment.[a]	3.86	3.77	-0.09	34	0.37	.356
S7: I would imagine that most students would learn to participate in/use the face-to-face/Aivaluate assessment very quickly	3.63	4.06	0.43	34	–2.00	.027*
S8: I found the face-to-face/Aivaluate assessment very cumbersome.[a]	3.51	3.97	0.46	34	–1.96	.029*
S9: I felt very confident during the face-to-face/Aivaluate assessment	3.40	3.89	0.49	34	–1.84	.037*
S10: I needed to learn a lot of things before I could get going with the face-to-face/Aivaluate assessment.[a]	3.60	3.86	0.26	34	–0.90	.187
Students' experience (additional items)						
A1: The face-to-face/Aivaluate assessment was a positive experience	4.17	3.83	-0.34	34	1.34	.095
A2: I felt comfortable expressing my thoughts during the face-to-face/Aivaluate assessment	4.00	3.89	-0.11	34	0.42	.340
A3: The face-to-face/Aivaluate assessment helped me to better understand my work	3.89	3.31	-0.57	34	2.77	.004*
A4: I did not feel anxious during the face-to-face/Aivaluate assessment	2.66	3.91	1.26	34	-4.46	< .001*

A1–A4 bespoke items presented in addition to the 10 standard SUS items to evaluate students' experience.

[a] Responses of negative items were reversed to align with positive items, higher scores indicate positive responses.

* Indicates statistically significant < = 0.05 confidence level.

significant differences in four out of the ten SUS items between face-to-face and AIvaluate assessments. For item S4, participants report significantly higher scores for the AIvaluate format (M=3.97, SD=1.06) compared to face-to-face (M=3.54, SD=1.10), $t(34)$=-1.74, p=.046. This suggests that AIvaluate assessments are perceived as requiring less external support, indicating greater accessibility. Similarly, item S7 shows a significant preference for the AIvaluate format (M=4.06, SD=1.04) compared to face-to-face (M=3.63, SD=1.04), $t(34)$=-2.00, p=.027. This highlights the perceived efficiency of learning to use the AIvaluate platform. Item S8 also shows significant results, with participants rating AIvaluate higher (M=3.97, SD=1.16) than face-to-face (M=3.51, SD=1.05), $t(34)$=-1.96, p=.029, suggesting that AIvaluate was seen as less cumbersome. Finally, item S9 was significantly higher for AIvaluate (M=3.89, SD=0.78) compared to face-to-face (M=3.40, SD=1.10), $t(34)$=-1.84, p=.037, reflecting increased confidence in the virtual assessment format. For the remaining six SUS items, no statistically significant differences are observed. For example, item S2 shows similar mean scores between face-to-face (M=3.86, SD=0.96) and AIvaluate (M=3.94, SD=0.89), $t(34)$=-0.38, p=.353, indicating that neither assessment format is perceived to take predominance.

The additional items (A1–A4), designed to evaluate participants' subjective experience with the assessments, also reveal significant findings. For item A3, AIvaluate (M=3.31, SD=1.01) scores significantly lower than face-to-face (M=3.89, SD=0.89), $t(34)$=2.94, p=.004, indicating that students' have a stronger sense of understanding of their work through face-to-face assessments. Item A4 shows a notable difference, with AIvaluate (M=3.91, SD=0.77) rated significantly higher than face-to-face (M=2.66, SD=1.35), $t(34)$=-4.46, p<0.001. This indicates that participants experience less anxiety during AIvaluate assessments. For items A1-2, no statistically significant differences are found, with both formats receiving comparable ratings.

Student preferences

The third research question aims to identify students' preferences for AIvaluate assessments over traditional face-to-face assessments, or vice versa. As explained in Sect. "Qualitative open-response survey data analysis (RQ3)", reflexive thematic analysis is conducted on the open-survey response data collected. The thematic analysis of the interviews reveals a number of high-level themes and associated sub-themes as illustrated in Table 4.

To illustrate the contrasting, and sometimes contradictory, student experiences across assessment formats, Table 5 presents a comparative thematic overview including verbatim quotes from the original dataset. The table illustrates how the same assessment dimensions were experienced as strengths across both formats, albeit in contrasting ways, thereby highlighting divergent student preferences rather than a single dominant mode. The quotes included in this table were intentionally selected to exemplify the breadth and variance of participant experiences within each theme and sub-theme. In contrast, the quotes used in the detailed thematic analysis were those most representative of the central meaning of each theme, capturing more typical rather than extreme perspectives.

The themes and sub-themes are now presented within the overarching template used to consider the results, which is the strengths and challenges of each respective mode of assessment. In order to maintain anonymity of research participants, any identifiable data is removed from participant quotes in this section.

Table 4 Thematic map of face-to-face and AIvaluate themes and sub-themes

Assessment	Format	Themes	NP*	Sub-Themes	NP*
Performance based Assessments (Viva voce)	Face-to-face	Strengths of human-led assessments	19	Interaction and engagement	15
				Non-verbal cues	12
				Natural conversation	8
		Challenges of human-led assessments	17	Increased anxiety	9
				Familiar teachers	16
	AIvaluate (virtual)	Strengths of technology-led assessments	21	Reduced pressure	14
				Response times	14
				Ease of use	11
				Feedback and reflection	15
		Challenges of technology-led assessments	18	Technical issues	8
				Impersonal interaction	10
				Waiting times	9

*NP = Number of Participants that made reference to the theme or sub-theme

AIvaluate: strengths of technology-led assessments

The following sub-themes illustrate the perceived strengths of the AIvaluate platform, as described in participant reflections on assessment comfort, control over responses, usability and opportunities for reflection. These sub-themes are clustered into an overarching theme that emphasises how the AI-based viva format reduces pressure, encourages thoughtful responses and fosters deeper engagement with learning.

Reduced pressure This sub-theme captures how the AI-based format helps students feel less anxious and less judged compared to face-to-face assessments. Codes here consistently reflect reduced intimidation, less feeling of judgement from teachers, increased comfort and a stronger sense of safety when students interact with the system rather than a teacher. Participants suggest that AIvaluate provides a relaxed and less intimidating environment compared to face-to-face assessments. By removing the immediate presence of a teacher, students feel less judged and more comfortable expressing themselves, underlining how comparatively at ease they felt during the virtual format of assessment. Similarly, it is suggested that there was less pressure when explaining concepts to an AI agent, suggesting that the virtual platform alleviates some nervousness associated with being directly observed by a person.

> *"I felt like there was a lot more pressure on me in the face-to-face meeting"* (Participant 8)
>
> *"I didn't have to feel the pressure of explaining in front of the teacher"*
>
> (Participant 35)

Furthermore, the lack of visible teacher reactions is also seen as contributing to a sense of safety from the student's perspective.

Table 5 Comparative thematic contrasts between face-to-face and AIvaluate viva voce assessments

Sub-theme	AI-preferred experience	Human-preferred experience	Interpretive insight
Interaction and engagement	*"I was much more calm when not facing face to face with a tutor, I was not as nervous and I was able to form better sentences before sending a reply"*	*"The conversation felt more genuine and it led me to be more actively engaged"* *"I genuinely enjoyed the face-to-face question and answer session"*	AIvaluate supported calm, structured responses; face-to-face enhanced engagement and authenticity
Non-verbal cues	*"I liked that I was not in front of a real person so I did not feel intimidated"* *"Not having to see a face expecting a good ans"*	*"Being able to see the teacher's expressions while having the conversation helped me figure out which aspect of the conversation was interesting"*	AIvaluate reduced perceived judgement; face-to-face cues supported reassurance and understanding
Natural conversation	*"Having the option to write, I am able to express myself more fully"* *"It felt more open to answer clearly and think clearly"*	*"I enjoy organic and spontaneous conversations, which would allow me to engage with the person I am talking to"* *"It felt more 'human' and natural"*	AIvaluate enabled reflective expression; face-to-face supported spontaneous dialogue
Increased anxiety	*"I felt like there was a lot more pressure on me in the face to face meeting"* *"I felt more relaxed"*	*"Even though I felt anxious at times it still made me feel like I gained perspective once I completed it"*	AIvaluate reduced anxiety; face-to-face promoted perspective despite pressure
Familiar teachers	*"I felt more confident with typing down the things I think instead of talking to a teacher that unwell known"*	*"I knew of my teacher, and I knew how her personality was, so it was easy for me to just explain things"* *"My teacher was nice and outgoing"*	AIvaluate minimised discomfort with unfamiliar assessors; familiarity eased face-to-face communication
Reduced pressure	*"I liked that I didn't have to feel the pressure of explaining in front of the teacher"* *"it was less pressure to give instant responses"*	*"I could tell if the teacher's response was positive or not, which made me more comfortable"* *"I found it less formal than AI so it eased my stress"*	AIvaluate lowered immediate performance pressure; face-to-face feedback increased comfort
Response times	*"time to think without having a teacher right in front of us giving us the pressure to answer immediately"* *"The slight delay allowed time for me to reflect*	*"Communication was quicker and kept the flow going"* *"Responses to what I said was faster and more insightful"*	AI delays allowed reflection; human responses sustained conversational flow
Ease of use	*"It was a simple system and easy to use"* *"very easy to use and un stressful"*	*"allowed for more meaningful interactions and was more engaging"* *"easier to communicate with the teacher"*	AIvaluate was simple and low-stress; face-to-face felt more intuitive
Feedback and reflection	*"It gave me ideas on how to improve it and possible extensions"* *"quite helpful in having me reflect back"*	*"The teacher still understood me due to my body language"* *"asked very engaging questions that made me think"*	AIvaluate encouraged structured reflection; teachers enabled adaptive feedback
Technical issues	*"The dictation feature was great because I don't type very fast"*	*"enabled me to be able to fix my responses as I am saying it"* *"Communication was quicker"*	AIvaluate improved access but introduced friction; face-to-face allowed instant clarification

Table 5 (continued)

Sub-theme	AI-preferred experience	Human-preferred experience	Interpretive insight
Impersonal interaction	*"I felt more relaxed and i also think i was less nervous"* *"good option for people who are anti-social"*	*"I prefer talking to a real person because the human has emotions"* *"felt more 'human' talking to an actual person"*	AIvaluate reduced social strain; human interaction offered emotional connection
Waiting times	*"The slight delay allowed time for me to reflect and think about my response"* *"Talking to the AI allowed for more time to think"*	*"Face to face meetings enabled us to go into more detail freely"* *"It took less time"*	AIvaluate supported reflection; face-to-face prioritised efficiency

> *"There was less judgement because, frankly, I couldn't see the teacher's reactions to my response"*
>
> (Participant 16)

These findings suggest that the AIvaluate format offers a more comfortable and less-pressured assessment environment for students to articulate their thoughts during viva voce PBAs.

Response times This sub-theme reflects how participants value the perception of having more time to think, compose and refine answers during AIvaluate sessions. Codes frequently highlight that the absence of immediate pressure to respond, combined with the option to write and edit, gives students a greater sense of control and confidence. Although neither face-to-face nor AIvaluate viva voce sessions are timed, participants have the perception that they have more time to compose and refine answers when interacting with AIvaluate. Participants assume that there is an allowance for deliberate thought and editing when engaging with a virtual system. This contrasts sharply with face-to-face sessions during which participants assume that there is a need to respond immediately to questions posed by the examiner. The opportunity to take time to reflect before responding to AIvaluate provides students with a sense of control and confidence. They also feel that they have the added capacity to deliver more thoughtful, considered, and structured answers, which is particularly valued by participants. Furthermore, some participants suggest that they prefer to articulate their ideas in writing.

> *"typing out my responses was a great feature because I was able to proofread each response I write and ensure that it is conveying what I wanted it to"*
>
> (Participant 5).
>
> *"AIvaluate allows us time to think without having a teacher right in front of us giving us the pressure to answer immediately"*
>
> (Participant 16).

"having the option to write, I am able to express myself more fully and allowing me to edit and add so that I am saying everything that I want to say"

(Participant 26).

Ease of use This sub-theme synthesises codes describing the accessibility, flexibility and simplicity of the AIvaluate platform. Participants note that its user-friendly design and choice of input method (typing or speaking) cater for diverse communication preferences, particularly supporting those who are less confident with verbal interaction. Participants tend to report that AIvaluate is user-friendly and adaptable, with participants appreciating its straightforward design and the flexibility to choose between typing or speaking their responses. This suggests that the platform caters for a range of communication preferences. The simplicity of the system also resonates with users. Moreover, the option to type answers is particularly beneficial for those who find verbal communication challenging.

"I think AIvaluate would be good for people who are not comfortable with talking face to face"

(Participant 33)

"It was a simple system and easy to use"

(Participant 21)

"We could type in our answers rather than voicing them out. This is good for people who have difficulties in explaining or expressing their thoughts"

(Participant 11)

Feedback and reflection This sub-theme captures how AIvaluate prompts critical thinking and deeper engagement. Codes emphasise the ability of the system to generate subject-specific questions, offer actionable suggestions and encourage students to reflect on the quality of their work, thereby extending beyond traditional assessment towards intellectual development. AIvaluate is seen as having the ability to prompt critical thinking and self-reflection, which is seen as another strength highlighted by participants. The platform's questions are perceived as insightful and subject specific, encouraging students to engage more deeply with their work. AIvaluate is able to generate actionable suggestions for improvement to participants, which resonates well with them.

"Responses were somewhat insightful at times when talking about my extended essay; it did ask me questions regarding my EE topic which made me critically think about the quality of the writing"

(Participant 2)

"The AI was great in analysing my EE from a simple prompt. It gave me ideas on how to improve it and possible extensions"

(Participant 31)

The detailed nature of the questions presented to participants about their work not only tests their understanding, but also provides clarity on areas where they can improve the answers they give during the viva voce. It appears that AIvaluate transcends the traditional assessment function by acting as a tool for intellectual development and deeper engagement.

"I liked how it told me that I answered the question wrongly and provided a question for clarification to better understand what it was looking for"

(Participant 22)

AIvaluate: challenges of technology-led assessments

Despite its strengths, participants also identify a number of challenges when engaging with AIvaluate. The following sub-themes synthesise reflections on technical barriers, lack of human connection and waiting times, highlighting tensions between efficiency and authentic interaction.

Technical issues This sub-theme describes how technical difficulties, such as system delays, hardware problems and dictation errors, disrupt the flow of assessments. Codes consistently associate these disruptions with frustration and a less smooth experience overall. While AIvaluate is generally well-received, technical difficulties present significant challenges for some participants. Delays in response times, unresponsive keyboards, and dictation errors disrupt the flow of the assessment.

"AIvaluate took long to respond which meant I had to wait for about a few minutes before getting the next response"

(Participant 1)

"the dictation feature cut out and didn't really translate my answer well, so I had to type my responses"

(Participant 35)

In terms of the keyboard, some students note low-quality hardware which makes the experience of using AIvaluate difficult at times.

"the main thing that disrupted my experience was the keyboard as it was quite hard to type"

(Participant 12)

Impersonal interaction This sub-theme reflects participants' concerns about the lack of warmth and emotional connection during AIvaluate sessions. Codes grouped here emphasise feelings of disengagement, robotic tone and the absence of authentic human interaction. Several participants describe AIvaluate as lacking the human touch, which makes the interaction feel less personal and unengaging. This appears to be further compounded by the robotic tone of the text-to-speech feature.

"to put it bluntly, the experience with AIvaluate system was quite boring and unengaging"

(Participant 4)

"I prefer talking to a real person because the human has emotions"

(Participant 15)

"... the voice, because it was very robotic, made it harder for me to take it seriously"

(Participant 28)

Waiting times This sub-theme highlights how students are negatively affected by the variability of response speeds, whether due to system processing or teacher selection of replies. Codes frequently reference delays as breaking conversational flow and causing frustration. Delays in receiving responses from the teacher are another point of frustration for participants, as prolonged waiting disrupts the natural flow of the assessment, emphasising the variability and unpredictability of system responses. Teachers sometimes take time to think about the suggested responses or form their own, in which case the student is waiting.

"the response time sometimes could be quite long depending on the question asked/answer given"

(Participant 3)

"responses were slower than I thought a chatbot would reply"

(Participant 22)

This indicates a mismatch between user expectations and the system's performance. While not as frequently cited as other issues, these waiting times detract from the overall efficiency of the platform and occasionally disrupt the students' focus during the assessment process.

Face-to-face: strengths of human-led assessments

Participants reflect positively on several features of traditional face-to-face assessments, particularly their interactivity, immediacy and authenticity. The following sub-themes demonstrate how in-person assessments allow for richer dialogue, use of non-verbal cues and spontaneous conversation.

Interaction and engagement This sub-theme captures how the immediacy and presence of a teacher creates more dynamic and interactive dialogue. Codes highlight spontaneity, responsiveness and stronger engagement compared with AI-mediated sessions. Face-to-face assessments are praised for their ability to facilitate meaningful interaction and deeper engagement. Participants note that the physical presence of a teacher helps to foster a natural flow to conversations, enabling them to be more dynamic and make immediate adjustments to their responses. This facilitates a more interactive dialogue that feels fluid and responsive compared with the equivalent virtual setting. The dynamic nature of face-to-face interactions also allows students to adjust their answers in real time. Furthermore, this more dynamic interaction enables some participants to feel more expressive and interactive in their responses.

> *"face to face, it was more engaging and flowed naturally"*
>
> (Participant 10)
>
> *"face-to-face meetings enabled me to fix my responses as I am saying it"*
>
> (Participant 16).

Non-verbal cues This sub-theme reflects how students value the ability to read and respond to body language, facial expressions and other non-verbal signals. Codes suggest that such cues boost confidence, reassure students and improve conversational flow. Another advantage of face-to-face assessments is the ability to interpret and respond to non-verbal cues, which many participants find to be a useful part of communication. Non-verbal feedback, such as facial expressions and body language, helps students gauge whether their responses are understood or require clarification.

> *"Being able to see the teacher's expressions while having the conversation helped me figure out which aspect of the conversation was interesting."*
>
> (Participant 16)

The real-time feedback allows participants to adjust their responses on the fly, which they feel possibly improves the coherence and relevance of their answers. Additionally, non-verbal cues contribute to students' sense of confidence and comfort.

> *"I could tell if the teacher's response was positive or not, which made me more comfortable"*
>
> (Participant 14)

Natural conversation This sub-theme synthesises comments on the authenticity and spontaneity of in-person dialogue. Codes emphasise how unstructured, organic exchanges help students feel engaged and connected. The natural conversational element of face-to-face assessments is another aspect participants value. Some describe these interactions as organic, spontaneous and genuine, allowing for a natural flow that feels less structured and more authentic. Face-to-face interactions also enable a more fluid exchange of ideas, where participants can expand on their responses without feeling constrained by a rigid structure.

> *"the conversation felt more genuine and it led me to be more actively engaged"*
>
> (Participant 2)
>
> *"I enjoy organic and spontaneous conversations, which would allow me to engage with the person I am talking to"*
>
> (Participant 7)
>
> *"[the face-to-face assessment was] more 'human' and natural"*
>
> (Participant 27)

Face-to-face: challenges of human-led assessments

Alongside the benefits, participants also report challenges associated with face-to-face assessments. The following sub-themes capture how live, high-stakes interactions sometimes increase anxiety, create discomfort with unfamiliar teachers, and make assessments unpredictable.

Increased anxiety This sub-theme captures how the immediacy of in-person dialogue heightens pressure for some students, often leading to nervousness and difficulty articulating thoughts. Codes frequently reference performance anxiety and the stress of real-time responses. Contrary to some of the strengths, numerous participants report that face-to-face assessments induce heightened anxiety, particularly when they feel under pressure to deliver immediate and articulate responses. The physical presence of a teacher is sometimes perceived as intimidating, leading to nervousness and, in some cases, a struggle to communicate effectively.

> *"I actually felt quite a bit of pressure when explaining to my teacher which made me almost stumble over my words"*
>
> (Participant 35)

This anxiety is particularly pronounced when students feel unprepared or unsure about their answers, as the immediacy of the interaction leaves little room for reflection or revision. The perceived pressure to respond quickly in real-time conversations also exacerbates this anxiety.

> *"I felt that I had to answer much quicker as it was a conversation and the person was in front of me, maybe because of this I was more nervous"*
>
> (Participant 9)

This sense of urgency often prevents students from fully formulating their thoughts, potentially leading to responses that do not accurately reflect their knowledge or understanding. For students who are naturally more reserved or prone to social anxiety, this aspect of face-to-face assessments presents a significant challenge.

Familiar teachers This sub-theme reflects how interacting with unfamiliar teachers creates additional stress for students. Codes point to lack of rapport, reduced openness and increased self-consciousness as barriers to communication. Another challenging aspect of face-to-face assessments is the discomfort associated with unfamiliar teachers. Many participants express that working with a teacher they do not know well makes the experience more stressful and less conducive to open communication. This highlights how unfamiliarity creates a barrier to effective interaction. This discomfort is further compounded by a lack of established rapport, which some participants feel prevents them from expressing themselves freely.

> *"talking face to face with a teacher that I didn't know was also more stressful"*
>
> (Participant 28)
>
> *"it was nerve-wracking to have another teacher who did not know about my EE interviewing me"*
>
> (Participant 16)

This suggests that a lack of familiarity not only increases stress but also hinders the overall quality of the assessment. For some students, this unfamiliarity introduces an additional layer of anxiety, making it harder for them to feel at ease and confident in their performance. This theme highlights the importance of trust and rapport in face-to-face assessments and suggests that these factors can significantly impact students' experiences and outcomes.

Discussion

The findings of this study provided valuable insights into the effectiveness and acceptability of using AIvaluate, an emotionally intelligent LLM-augmented pedagogical conversational agent, to reduce student anxiety during oral PBAs. This section contextualises the results within the broader educational literature, discusses the implications for practice, and identifies areas for further research.

The primary aim of this study was to determine whether AIvaluate could reduce student anxiety levels during PBAs compared to traditional face-to-face assessments. The findings relating to *RQ1* revealed that, in terms of strengths of AIvaluate, students reported significantly lower anxiety levels during AIvaluate assessments. These findings aligned with previous research highlighting the potential of AI conversational agents to create a low-pressure environment conducive to student well-being (Klos et al., 2021; Liu et al., 2022). As a result of the open response surveys, some key factors that may have contributed to the observed reduced anxiety in AIvaluate assessments included the absence of direct teacher observation and participants having the perception that they had more time to respond to questions compared with the virtual setting. Students appreciated having the flexibility to compose and refine their answers in a virtual setting, which contrasted with the perceived need to respond immediately in face-to-face interactions. These results corroborated findings from (Alcorn & Cheesman, 2022), who emphasised the benefits of controlled environments in technology-assisted PBAs. Although reduced anxiety was observed in AI-mediated assessments, this should not be assumed to directly translate into improved learning outcomes. As discussed in Sect. "A discourse on student anxiety", moderate levels of anxiety could be facilitative for learning and performance; however, for some learners heightened anxiety may more readily exceed optimal arousal levels, increasing the likelihood of threat-based responses that undermine cognitive processing and assessment performance. Future research should therefore examine whether AI-mediated PBAs balance the reduction of debilitating anxiety with the maintenance of constructive challenge. In terms of challenges, accurate replication of human-like engagement was a limitation of AIvaluate, as some students noted the lack of emotional connection and spontaneity when using the application. While the AI conversational agent effectively reduced anxiety for the majority of participants, it is worth noting that some students preferred the more dynamic interaction that face-to-face assessments offered. This suggests that while technology-assisted assessments conducted through AI conversational agents, such as AIvaluate, might serve as a supportive alternative, they might not serve as a complete replacement for traditional methods, particularly for students who thrive on direct human interaction. Furthermore, it is important to note that AIvaluate did not capture discrete categories of anxiety (e.g. cognitive, somatic, social-evaluative), as the 1–10 slider in the self-reporting tool captured students' overall subjective anxiety intensity during assessment tasks. Given that performance-based oral assessments are particularly associated with cognitive worry and social-evaluative pressure, it is reasonable to interpret these ratings as reflecting assessment-related anxiety within those domains, rather than general or clinical anxiety. Moreover, the anxiety-related findings in this study should be understood in light of the underlying objectives of the EE viva voce, as described in Sect. "Assessment Objectives and Performance Measures". This assessment requires students to draw on multiple skill domains simultaneously; articulating conceptual understanding of their research, engaging in critical reflection, structuring coherent and well-organised responses, and communicating effectively in an oral format. These demands involve both cognitive and communicative complexity, which can naturally elevate anxiety levels, particularly when responses must be produced spontaneously in the presence of an examiner. While the present study focused on emotional and usability outcomes, it did not examine the relationship between these objectives and detailed performance metrics, which limits the extent to which anxiety levels can be interpreted alongside demonstrated proficiency in each domain. Future research could incorporate validated performance measures to explore how specific objective domains, such as oral fluency or critical reasoning, interact with different delivery formats to influence both emotional and performance outcomes.

The findings for *RQ2* revealed that AIvaluate demonstrated strong usability, with an overall SUS score indicating "good" usability, surpassing the score for face-to-face assessments. Significant differences were observed in areas such as perceived confidence and the need for external support, with AIvaluate scoring higher in both cases. This reflects the user-friendly design of the system and its ability to provide a supportive, less intimidating assessment environment. The thematic analysis highlighted specific advantages, such as reduced pressure, flexible response options, and actionable feedback, which enhanced the user experience. These attributes are consistent with the principles of the Universal Design for Learning educational framework, which emphasise flexibility and personalisation to accommodate diverse learner needs (Altowairiki, 2024; Meyer et al., 2014). However, technical issues including response delays and errors in speech-to-text functionality detracted from the overall experience for some students. These findings align with prior research on the limitations of AI conversational agent systems, emphasising the need for robust technical infrastructure and responsive design to ensure seamless user interaction (Padmasiri et al., 2023; Tyagi, 2024). Addressing these challenges is crucial for improving the reliability and scalability of AI conversational agent-based assessment systems.

The findings for *RQ3* revealed a range of preferences for, and against, the two assessment formats. While AIvaluate excelled in reducing anxiety, and providing a structured environment, face-to-face assessments were preferred for their natural conversation flow, real-time feedback, and engagement through non-verbal cues. These results highlight the complementary strengths of both formats. Students valued the interactive and engaging nature of face-to-face assessments, which facilitated adaptive dialogue and immediate clarification of concepts. However, the face-to-face format also seemed to elicit reports of higher anxiety, particularly in the presence of unfamiliar teachers. This finding is consistent with existing literature on social evaluation anxiety in educational settings (Mills et al., 2016). On the other hand, AIvaluate offered a less intimidating environment by removing the immediacy of direct observation, fostering comfort and enabling thoughtful responses, which could potentially better support learners who experience heightened anxiety that exceeds the individual zone of optimal functioning during PBAs. Nevertheless, the lack of human interaction and occasional technical glitches limited its ability to fully replicate the richness of face-to-face dialogue. These compromises highlight the potential for exploring the

Table 6 Future recommendations

#	Research recommendation	Source
RR1	Integrate AI-based assessment into range of assessment strategies	ESA, ORSA: RP
RR2	Explore the value of hybrid assessment models	ORSA: RP, EU, IE
RR3	Focus on developing more anthropomorphic agents	ORSA: II
RR4	Explore potential within formative and summative assessments	ESA, ORSA: RP, EU

RQ1 – Emotional State Analysis (ESA).

RQ2 – Systems Usability Scale Analysis (SUSA).

RQ3 – Open Response Survey Analysis (ORSA): Ease of Use (EU), Impersonal Interaction (II), Interaction and Engagement (IE), Reduced Pressure (RP).

development of hybrid assessment models that integrate the strengths of both face-to-face and virtual assessment formats.

In light of the key findings of this study, several recommendations for future research have been identified for the deployment of an AI conversational agents in PBAs. These research recommendations are presented in Table 6.

RR1. **Integrate AI-based assessment into range of assessment strategies:** As AIvaluate achieved promising results in reducing student anxiety, institutions should explore ways of integrating AI-based assessments into a range of existing assessment strategies to maximise its effectiveness and realise potential benefits. Candidate assessment strategies could include structured interviews, scenario-based questioning, and adaptive questioning techniques that respond to student input dynamically. By tailoring the assessment approach to the learning objectives, educators can leverage AI's versatility while ensuring that evaluations remain rigorous and aligned with curriculum goals.

RR2. **Explore the value of hybrid assessment models:** A blended approach that combines face-to-face with augmented AI-based assessment tools could help to address diverse student preferences and learning styles. A hybrid model of assessment could provide flexibility, allowing students to benefit from the personalised feedback of face-to-face human interactions whilst also benefitting from the reduced anxiety that AI-based applications like AIvaluate can offer. Such a model could help facilitate inclusivity and help educators better assess students' abilities across various contexts.

RR3. **Focus on developing more anthropomorphic agents:** To improve user experience and engagement, further research and development of AI-based assessment systems should focus on making interactions more human-like. Enhancing the emotional intelligence of AIvaluate, such as its ability to recognise and respond empathetically to student emotions, can help mitigate concerns about impersonal interactions. Anthropomorphic features, like adaptive language and tone, could also increase trust and acceptance among users.

RR4. **Explore potential within formative and summative assessments:** The potential for AI-based integrations in both formative and summative assessments should be further explored. AIvaluate can be utilised to provide immediate, personalised feedback during formative assessments, enabling students to track their progress and identify areas for improvement in real time. For summative assessments, AI systems can ensure consistency, reduce bias, and provide robust data analysis to support final grading. Developing innovative AI-based assessment formats that align with institutional learning objectives could revolutionize how student learning is evaluated.

Study limitations and future research

This study examined student experience and anxiety rather than performance outcomes. AIvaluate was associated with reduced anxiety. This may be especially relevant for learners whose anxiety exceeds their individual zone of optimal functioning in performance-based assessments, including those with social anxiety or neurodivergent profiles; however, reduced anxiety should not be assumed to benefit all students. It is important to note that this study did not explicitly recruit participants with identified social anxiety or neurodivergent profiles, nor did it collect data relating to these characteristics. As a result, it is not possible to determine whether AIvaluate is empirically beneficial for these learner groups, representing a limitation of the present study. Nevertheless, the observed reductions in anxiety suggest potential value for learners who experience anxiety that exceeds their individual zone of optimal performance during performance-based assessments, an implication that warrants targeted investigation in future research. Further investigation is required into measuring student performances and outcomes when using AI-powered assessment tools. Furthermore, this study and the associated results have focused on a specific demographic of students, which may limit generalisability, hence, future research should examine the applicability of AIvaluate across diverse educational contexts and age groups.

The study also assessed immediate emotional and usability outcomes. Longitudinal studies are needed to evaluate the long-term impact of AI conversational agent-based assessments on student learning and performance. Future research should focus on collecting, analysing and optimising system performance data and exploring the integration of advanced multimodal interaction features to address technical shortcomings. Future work may use multidimensional instruments to disaggregate types of anxiety; however, this study prioritised capturing immediate, intuitive self-reports that reflect how students experience anxiety dynamically during assessments. AIvaluate represents a promising advancement in educational technology, offering an emotionally intelligent, LLM-augmented conversational agent that reduces student anxiety and provides a more positive assessment experience, however, this solely in the context of student anxiety. While face-to-face assessments remain valuable for their dynamic interaction and engagement, AI-driven assessments may offer a viable alternative in specific contexts and for particular learner groups. By addressing technical challenges and integrating hybrid models, institutions can create more equitable and effective assessment frameworks that cater to the diverse needs of students.

Conclusion

This study explored the effectiveness of AIvaluate, an emotionally intelligent LLM-augmented pedagogical AI conversational agent, in reducing student anxiety during oral PBAs. The findings demonstrate that AIvaluate significantly reduces student anxiety levels compared to traditional face-to-face assessments while maintaining strong usability and potential for user acceptance, however, further investigation is necessary to ascertain whether reduced anxiety is beneficial for the students learning. AIvaluate has the potential to function as a facilitative mechanism for learners who experience heightened anxiety in performance-based assessments that exceeds optimal arousal levels. Such anxiety can increase the risk of threat-based responses and mask underlying competence. Key benefits of AIvaluate include its ability to create a low-pressure environment, facilitate thoughtful and structured responses, and provide flexible communication options. These attributes were instrumental in fostering a more inclusive and comfortable assessment experience, particularly for students prone to performance anxiety. The system design and integration of emotional intelligence through real-time self-reporting and adaptive responses contributed to its overall positive reception, as reflected in higher SUS scores compared to face-to-face assessments. However, limitations were noted, including occasional technical issues and the lack of human engagement that some students value in traditional assessment formats. Face-to-face assessments were praised for their interactive and conversational dynamics, immediate feedback, and the richness of non-verbal communication, which foster deeper engagement and natural dialogue. These advantages highlight the complementary strengths of face-to-face and AI conversational agent-based assessments. The findings suggest that AIvaluate can serve as an effective tool for reducing student anxiety in oral PBAs, offering a viable alternative to more traditional assessment formats. While this study focused on a specific demographic, its findings highlight the potential for AI conversational agents to transform assessments by reducing anxiety and promoting inclusivity. Future research should explore the application of such systems across diverse educational settings and evaluate their long-term impact on learning outcomes, including the comparison and analysis of performance scores in both traditional and AI led assessments. The development of hybrid assessment frameworks, integrating both AI and human elements, represents a promising avenue for creating adaptive, student-centered assessment practices that cater to a variety of needs and preferences. AIvaluate thus exemplifies the potential of innovative technology to advance educational equity and effectiveness in the modern classroom.

Acknowledgements The authors would like to thank all of the participants that took part in this study and Brunel University for providing the facilities and materials necessary to carry out the interactive sessions with participants.

Author's contributions AM, HY and DDZ conceived the study. HY and AM planned the study and formulated the experiment design. HY developed the application. HY performed the user trials. AM and HY were responsible for data analysis and writing the manuscript. HY, AM and DDZ reviewed and edited subsequent drafts of the manuscript.

Funding No funding was obtained for this study.

Data availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare that they have no competing interests.

Consent for publication Not applicable.

Ethics approval and consent to participate This research was approved by Brunel University of London Research Ethics Committee. Informed consent to participate was obtained from each participant through signing of informed consent forms. Participants were informed of their right to withdraw from the study at their own will without any repercussions at any point during or after the study.

References

Abbasian, M., Yang, Z., Khatibi, E., Zhang, P., Nagesh, N., Azimi, I., Jain, R., & Rahmani, A. M. (2024). Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients. *arXiv preprint arXiv:2402.10153*. https://doi.org/10.48550/arXiv.2402.10153

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics, 132*, Article 103978.

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, Article 100006.

Alcorn, S. R., & Cheesman, M. J. (2022). Technology-assisted viva voce exams: A novel approach aimed at addressing student anxiety and assessor burden in oral assessment. *Currents in Pharmacy Teaching and Learning, 14*(5), 664–670.

Alser, M., & Waisberg, E. (2023). Concerns with the usage of ChatGPT in academia and medicine: A viewpoint. *American Journal of Medicine Open, 9*, Article 100036.

Alshumaimeri, Y. A., & Alshememry, A. K. (2024). The extent of AI applications in EFL learning and teaching. *IEEE Transactions on Learning Technologies, 17*, 653–663. https://doi.org/10.1109/tlt.2023.3322128

Altowairiki, N. (2024). Blended learning and the universal design for learning: Practical implications and students' perceptions. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2024.2416496

Andrade, M. S. (2014). Dialogue and structure: Enabling learner self-regulation in technology-enhanced learning environments. *European Educational Research Journal, 13*(5), 563–574.

Angelidis, A., Solis, E., Lautenbach, F., van der Does, W., & Putman, P. (2019). I'm going to fail! Acute cognitive performance anxiety increases threat-interference and impairs WM performance. *PLoS ONE, 14*(2), Article e0210824.

Arent, S. M., & Landers, D. M. (2003). Arousal, anxiety, and performance: A reexamination of the inverted-U hypothesis. *Research Quarterly for Exercise and Sport, 74*(4), 436–444.

Aubrey, S. (2022). The relationship between anxiety, enjoyment, and breakdown fluency during second language speaking tasks: An idiodynamic investigation. *Frontiers in Psychology, 13*, Article 968946.

Baker, E. L. (1997). Model-based performance assessment. *Theory into Practice, 36*(4), 247–254. https://doi.org/10.1080/00405849709543775

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *J. Usability Studies, 4*(3), 114–123.

Bielak, J. (2025). To what extent are foreign language anxiety and foreign language enjoyment related to L2 fluency? An investigation of task-specific emotions and breakdown and speed fluency in an oral task. *Language Teaching Research, 29*(3), 911–941.

Bos, A. E., Huijding, J., Muris, P., Vogel, L. R., & Biesheuvel, J. (2010). Global, contingent and implicit self-esteem and psychopathological symptoms in adolescents. *Personality and Individual Differences, 48*(3), 311–316.

Braun, V., & Clarke, V. (2013). Successful qualitative research: A practical guide for beginners.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2012). *Thematic analysis*. American Psychological Association.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Brooke, J. (1996). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry., 3*, 34.

Brooke, J. (2013). SUS: A retrospective. *J. Usability Studies, 8*(2), 29–40.
Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods, 17*(4), 600.
Burić, I., & Frenzel, A. C. (2019). Teacher anger: New empirical insights using a multi-method approach. *Teaching and Teacher Education, 86*, Article 102895.
Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity, 56*(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y
Chamorro-Premuzic, T., Furnham, A., Dissou, G., & Heaven, P. (2005). Personality and preference for academic assessment: A study with Australian university students. *Learning and Individual Differences, 15*(4), 247–256. https://doi.org/10.1016/j.lindif.2005.02.002
Cherakara, N., Varghese, F., Shabana, S., Nelson, N., Karukayil, A., Kulothungan, R., Farhan, M. A., Nesset, B., Moujahid, M., & Dinkar, T. (2023). Furchat: An embodied conversational agent using llms, combining open and closed-domain dialogue with facial expressions. *arXiv preprint arXiv:2308.15214*. https://doi.org/10.48550/arXiv.2308.15214
Chin, E. C., Williams, M. W., Taylor, J. E., & Harvey, S. T. (2017). The influence of negative affect on test anxiety and academic performance: An examination of the tripartite model of emotions. *Learning and Individual Differences, 54*, 1–8.
Cho, E., Ju, U., Kim, E. H., Lee, M., Lee, G., & Compton, D. L. (2023). Relations among motivation, executive functions, and reading comprehension: Do they differ for students with and without reading difficulties? *Scientific Studies of Reading, 27*(4), 289–310.
Cotton, D., Cotton, P., & Shipway, J. (2023). Chatting and cheating. Ensuring academic integrity in the era of ChatGPT. EdArXiv. *Preprint posted online January*, *10*.
D'Amore, S., Maurisse, A., Gubello, A., & Carone, N. (2023). Stress and resilience experiences during the transition to parenthood among Belgian lesbian mothers through donor insemination. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph20042800
Danieli, M., Ciulli, T., Mousavi, S. M., Silvestri, G., Barbato, S., Di Natale, L., & Riccardi, G. (2022). Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: Randomized controlled trial. *JMIR Mental Health, 9*(9), Article e38067.
Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64*(1), 5–31. https://doi.org/10.17763/haer.64.1.j57n353226536276
Davis, F. D. (1989). Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption, 205*(219), 5.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019) (2019–01–01). Proceedings of the 2019 Conference of the North.
Donate, Á. (2022). Task anxiety, cognition and performance on oral tasks in L2 Spanish. *Journal of Spanish Language Teaching, 9*(1), 1–18.
Dörnyei, Z. (2020). Innovations and Challenges in Language Learning Motivation. *2*. https://doi.org/10.4324/9780429485893
Doug, A., Fred, N., W, M., & V, R. (1988). *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School*. https://go.exlibris.link/zjTKDF7d
Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., & Ahuja, M. (2023). Opinion paper:"So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, Article 102642.
Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*(2), 350–383.
Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology, 13*, Article 100060.
El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems, 38*(3), Article e12667.
Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., & Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology, 56*(3), 1074–1091.
Ernst, J. V. (2008). Analysis of cognitive and performance assessments in an engineering/technical graphics curriculum. *Journal of sTEm Teacher Education, 45*, 7.
Eysenck, M. W. (1992). Anxiety: The cognitive perspective. *Erlbaum*. https://doi.org/10.4324/9780203775677
Eysenck, M. (2012). Attention and arousal: Cognition and performance. *Springer Science & Business Media., 3*, 34.
Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion (Washington, D.c.), 7*(2), 336.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), Article e7785.

Gallardo, K. (2020). Competency-based assessment and the use of performance-based evaluation rubrics in higher education: Challenges towards the next decade. *Problems of Education in the 21st Century, 78*(1), 61–79. https://doi.org/10.33225/pec/20.78.61

Gonda, D. E., & Chu, B. (2019) (2019–12–01). Chatbot as a learning resource? Creating conversational bots as a supplement for teaching assistant training course. *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*

Gonda, D. E., Luo, J., Wong, Y.-L., & Lei, C.-U. (2018), (2018–12–01). Evaluation of Developing Educational Chatbots Based on the Seven Principles for Good Teaching. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*

Hanoch, Y., & Vitouch, O. (2004). When less is more: Information, emotional arousal and the ecological reframing of the Yerkes-Dodson law. *Theory & Psychology, 14*(4), 427–452.

Hawanti, S., & Zubaydulloevna, K. M. (2023). AI chatbot-based learning: Alleviating students' anxiety in English writing classroom. *Bulletin of Social Informatics Theory and Application, 7*(2), 182–192.

Heydarnejad, T., Tagavipour, F., Patra, I., & Farid Khafaga, A. (2022). The impacts of performance-based assessment on reading comprehension achievement, academic motivation, foreign language anxiety, and students' self-efficacy. *Language Testing in Asia, 12*(1), 51.

Hungerford, C., Walter, G., & Cleary, M. (2015). Clinical case reports and the viva voce: A valuable assessment tool, but not without anxiety. *Clinical Case Reports, 3*(1), 1–2. https://doi.org/10.1002/ccr3.225

Imura, T. (2004). The effect of anxiety on oral communication skills. 2nd Annual International Conference on Cognition. *Language and Special Education Research*

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth, 6*(11), Article e12106.

Iqbal, I., Naqvi, S., Abeysundara, L., & Narula, A. (2010). The value of oral assessments: A review. *The Bulletin of the Royal College of Surgeons of England, 92*(7), 1–6. https://doi.org/10.1308/147363510x511030

Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology, 52*(6), 1122.

Jiang, J., Vauras, M., Volet, S., Salo, A.-E., & Kajamies, A. (2019). Autonomy-supportive and controlling teaching in the classroom: A video-based case study. *Education Sciences, 9*(3), 229.

Kabbar, E., & Barmada, B. (2024). Assessment Validity in the Era of Generative AI Tools.

Kalyuga, S. (2008). Managing cognitive load in adaptive multimedia learning. *IGI Global., 3*, 34.

Kearney, L. (2013). *The intra and interpersonal effects of observer and field perspective imagery in social anxiety* Kingston University].

Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial intelligence–based chatbot for anxiety and depression in university students: Pilot randomized controlled trial. *JMIR Formative Research, 5*(8), Article e20678.

Knight, R. A., Dipper, L., & Cruice, M. (2013). The use of video in addressing anxiety prior to viva voce exams. *British Journal of Educational Technology, 44*(6), E217–E219. https://doi.org/10.1111/bjet.12090

Krieglstein, F., Beege, M., Rey, G. D., Ginns, P., Krell, M., & Schneider, S. (2022). A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psychology Review, 34*(4), 2485–2541.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association, 25*(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Lee, Y.-F., Hwang, G.-J., & Chen, P.-Y. (2022). Impacts of an AI-based cha bot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educational Technology Research and Development, 70*(5), 1843–1865.

Lee, Y.-C., & Fu, W.-T. (2019) (2019–03–16). Supporting peer assessment in education with conversational agents. *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*

Li, E. Y., & Jan, A. (2023). Impact of artificial intelligence (AI) in enhancing productivity and reducing stress among students.

Lin, G. H. C. (2008). Pedagogies Proving Krashen's Theory of Affective Filter. *Online submission*.

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions, 27*, Article 100495.

Loveys, K., Fricchione, G., Kolappa, K., Sagar, M., & Broadbent, E. (2019). Reducing patient loneliness with artificial agents: Design insights from evolutionary neuropsychiatry. *Journal of Medical Internet Research, 21*(7), Article e13664.

Manole, A., Cârciumaru, R., Brînzaş, R., & Manole, F. (2024). Harnessing AI in anxiety management: A chatbot-based intervention for personalized mental health support. *Information, 15*(12), 768.

Maryadi, J. A., Santoso, H., & Isa, Y. K. (2017) (2017–11–01). Development of Personalized Pedagogical Agent for Student-Centered e- Learning Environment. *2017 7th World Engineering Education Forum (WEEF)*

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal Design for Learning: Theory and Practice*. CAST, Incorporated. https://books.google.com.bn/books?id=ZGt0oAEACAAJ

Miller, E. K., & Buschman, T. J. (2015). Working memory capacity: Limits on the bandwidth of cognition. *Daedalus, 144*(1), 112–122.

Mills, B., Carter, O., Rudd, C., Claxton, L., & O'Brien, R. (2016). An experimental investigation into the extent social evaluation anxiety impairs performance in simulation-based learning environments amongst final-year undergraduate nursing students. *Nurse Education Today, 45*, 9–15.

Navarro, E., Macnamara, B. N., Glucksberg, S., & Conway, A. R. (2020). What influences successful communication? An examination of cognitive load and individual differences. *Discourse Processes, 57*(10), 880–899.

Newell, S. J. (2023). Employing the interactive oral to mitigate threats to academic integrity from ChatGPT. *Scholarship of Teaching and Learning in Psychology., 3*, 435.

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122.

Padmasiri, P., Kalutharage, P., Jayawardhane, N., & Wickramarathne, J. (2023). AI-Driven User Experience Design: Exploring Innovations and Challenges in Delivering Tailored User Experiences. *2023 8th International Conference on Information Technology Research (ICITR)*

Pearce, J., & Chiavaroli, N. (2023). Rethinking assessment in response to generative artificial intelligence. *Medical Education, 57*(10), 889–891.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341.

Potter, B. S., Ernst, J. V., & Glennie, E. J. (2017). Performance-based assessment in the secondary STEM classroom: Performance-based assessments provide a vehicle to demonstrate student procedural knowledge as well as higher-order thinking abilities.(feature article). *Technology and Engineering Teacher, 76*(6), 18.

Proudfoot, K. (2023). Inductive/deductive hybrid thematic analysis in mixed methods research. *Journal of Mixed Methods Research, 17*(3), 308–326. https://doi.org/10.1177/15586898221126816

De la Puente, G., Silva, A., & Felix, R. (2024). Development of a Chatbot Powered by Artificial Intelligence to Diagnose and Improve Stress and Anxiety Levels in University Students. 2024 IEEE XXXI International Conference on Electronics, Electrical Engineering and Computing (INTERCON)

Rao, R. (2020). *To Study the Effectiveness of Homoeopathic Treatment Along with Counselling as Compared to Only Counselling for Test Anxiety Among Students* Rajiv Gandhi University of Health Sciences (India)].

Rapee, R. M., & Heimberg, R. G. (1997). A cognitive-behavioral model of anxiety in social phobia. *Behaviour Research and Therapy, 35*(8), 741–756.

Redifer, J. L., Bae, C. L., & DeBusk-Lane, M. (2019). Implicit theories, working memory, and cognitive load: Impacts on creative thinking. *SAGE Open, 9*(1), Article 2158244019835919.

Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist, 44*(3), 159–175.

Sapolsky, R. M. (2003). Stress and plasticity in the limbic system. *Neurochemical Research, 28*(11), 1735–1742.

Scholing, A., & Emmelkamp, P. M. (1990). Social phobia: Nature and treatment In Handbook of social and evaluation anxiety. *Springer., 3*, 269–324.

Scott, M., & Unsworth, J. (2018). Matching final assessment to employability: Developing a digital viva as an end of programme assessment. *Higher Education Pedagogies, 3*(1), 373–384. https://doi.org/10.1080/23752696.2018.1510294

Shafiee Rad, H., & Jafarpour, A. (2023). Effects of well-being, grit, emotion regulation, and resilience interventions on L2 learners' writing skills. *Reading & Writing Quarterly, 39*(3), 228–247. https://doi.org/10.1080/10573569.2022.2096517

Shorey, S., Ang, E., Yap, J., Ng, E. D., Lau, S. T., & Chui, C. K. (2019). A virtual counseling application using artificial intelligence for communication skills training in nursing education: Development study. *Journal of Medical Internet Research, 21*(10), Article e14658. https://doi.org/10.2196/14658

Skantze, G., & Hjalmarsson, A. (2010). Towards incremental speech generation in dialogue systems. *Proceedings of the SIGDIAL 2010 Conference*

Souppez, J.-B. R., Goswami, D., & Yuen, J. (2023). Assessment and Feedback in the Generative AI Era: Transformative Opportunities, Novel Assessment Strategies and Policies in Higher Education. *International Federation of National Teaching Fellows Symposathon 2023*

Soureshjani, K. H. (2013). A study on the effect of self-regulation and the degree of willingness to communicate on oral presentation performance of EFL learners. *International Journal of Language Learning and Applied Linguistics World, 4*(4), 166–177.

Stafford, D. K. C. (2024). IMPACT OF TRAUMA. *Skills for Safeguarding: A Guide to Preventing Abuse and Fostering Healing in the Church*.

Steinfeld, A., Jenkins, O. C., & Scassellati, B. (2009). The oz of wizard: simulating the human for interaction research. *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Tafazoli, D., & Meihami, H. (2023). Narrative inquiry for CALL teacher preparation programs amidst the COVID-19 pandemic: Language teachers' technological needs and suggestions. *Journal of Computers in Education, 10*(1), 163–187. https://doi.org/10.1007/s40692-022-00227-x

Tawfik, E., Ghallab, E., & Moustafa, A. (2023). A nurse versus a chatbot- The effect of an empowerment program on chemotherapy-related side effects and the self-care behaviors of women living with breast cancer: A randomized controlled trial. *BMC Nursing, 22*(1), 102.

Thamby, S. A., Muttusamy, B., Yee, S. M., Ayapanaido, T., & Parasuraman, S. (2016). Investigation of stressors affecting a sample of pharmacy students and the coping strategies employed using Modified Academic Stressors Scale and Brief Cope Scale: A prospective study. *Journal of Young Pharmacists, 8*(2), 122–127. https://doi.org/10.5530/jyp.2016.2.12

Tyagi, A. K. (2024). Applications of Human Computer Interaction, Explainable Artificial Intelligence and Conversational Artificial Intelligence in Real-life Sectors In Data Analytics and Artificial Intelligence for Predictive Maintenance in Smart Manufacturing. *CRC Press., 4*, 282–325.

Tzirides, A. O., Saini, A., Zapata, G., Searsmith, D., Cope, B., Kalantzis, M., Castro, V., Kourkoulou, T., Jones, J., & da Silva, R. A. (2023). Generative AI: Implications and applications for education. *arXiv preprint arXiv:2305.07605*.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 23.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.

Wang, Q., Jing, S., Camacho, I., Joyner, D., & Goel, A. (2020) (2020–04–25). Jill Watson SA: Design and Evaluation of a Virtual Agent to Build Communities Among Online Learners. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*

Wang, H., Tang, S., & Lei, C. U. (2024). AI Conversational Agent Design for Supporting Learning and Well-Being of University Students.

Weinberg, R. S., & Mellano, K. T. (2023). Arousal Management In Routledge Handbook of Applied Sport Psychology. *Routledge., 3*, 533–542.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

Xu, Y., Zhu, J., Wang, M., Qian, F., Yang, Y., & Zhang, J. (2024). The impact of a digital game-based AI chatbot on students' academic performance, higher-order thinking, and behavioral patterns in an information technology curriculum. *Applied Sciences, 14*(15), Article 6418.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459–482. https://doi.org/10.1002/cne.920180503

Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Pedagogical AI conversational agents in higher education: A conceptual framework and survey of the state of the art. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-025-10447-4

Habeeb Yusuf is an experienced academic, currently employed as a Course Leader and Senior Lecturer in Computer Science at University Academy 92', a newly established higher education institute based in Manchester setup in partnership with Lancaster University. Previously, he held different positions at the University of Bolton group, including Systems Developer and Senior Lecturer. Furthermore, Habeeb is currently a Doctoral Researcher in the Department of Computer Science at Brunel University London.

Dr Arthur Money is a Reader and Co-Director of Teaching and Learning in the Department of Computer Science at Brunel University London, where he also received his PhD in Multimedia Computing in 2007. Dr Money's research focuses on the user-centred design, development and evaluation of multimedia computing systems and the effective deployment of these systems with users who have complex needs spanning a range of domains including older adults, healthcare, education, and defence.

Dr Damon Daylamani-Zad is a Senior Lecturer in Digital Media at Brunel University. Damons main research activities are within the Creative Computing, Games and Digital Media domain, and are specifically focused on Applications of Artificial Intelligence in Games and Digital Media. He has worked on using Machine Learning for automated music generation, strategy planning, user modelling and game personalisation, and use of AI and gamification in education and training such as accessibility design, employability skills, and reading interventions in Dyslexia.

Authors and Affiliations

Habeeb Yusuf[1] · Arthur Money[1] · Damon Daylamani-Zad[1]

✉ Arthur Money
Arthur.Money@brunel.ac.uk

Habeeb Yusuf
Habeeb.Yusuf@brunel.ac.uk

Damon Daylamani-Zad
Damon.Daylamani-zad@brunel.ac.uk

1 Department of Computer Science, Brunel University of London, Kingston Lane, Uxbridge UB8 3PH, UK