

Learner Emotions and AI Scoring: Integrating Affective Signals in Language Evaluation

SAGE Open
April-June 2026: 1–13
© The Author(s) 2026
DOI: 10.1177/21582440261447928
journals.sagepub.com/home/sgo

Sabahattin Yeşilçınar¹

Abstract

The rapid evolution of artificial intelligence (AI) and the Internet of Things (IoT) is reshaping language assessment practices, yet most AI-driven approaches remain limited in their ability to recognize the affective dimensions that shape learner performance. This study explores how real-time emotional and physiological signals—such as heart rate variability and facial expressions—can be integrated into AI-mediated oral assessment to enhance fairness, validity, and learner support. Using a combination of generative AI (GPT-4), OpenAI Whisper, affective computing, and IoT-enabled biofeedback, an experimental platform was developed to capture and respond to learners' emotional states during speaking tasks. Grounded in Weir's sociocognitive validity framework and Shneiderman's human-centered AI principles, the study involved 24 intermediate English learners and examined the correlation between affective data and oral performance, the alignment between AI and human scoring, and learner perceptions of fairness and emotional support. Results showed that emotional states significantly impacted language output and that integrating affective input improved the alignment between AI and human ratings, particularly in fluency and coherence. Participants generally perceived the system as transparent and empathetic but raised concerns about emotional surveillance and data privacy. The findings contribute to emerging discussions on human-centered assessment design by highlighting how affective responsiveness in AI-supported assessment can foster interpretive fairness, reduce performance anxiety, and support learner agency.

Plain Language Summary

How AI Can Understand Learners' Emotions to Make Language Tests Fairer

This study looks at how learners' emotions, such as stress or calmness, can affect how well they perform during speaking exams in English. Traditional language tests and even new AI-based tests often ignore how nervous or anxious a learner feels, even though these emotions can impact their ability to speak clearly or fluently. To explore this, the researcher created a speaking test system using artificial intelligence (AI), speech recognition, and wearable devices that track things like heart rate and facial expressions. These tools helped the system notice how learners were feeling while they spoke. The study involved 24 university students and compared scores from the AI with scores from human teachers. The results showed that when the AI took emotions into account, it gave scores that were closer to what human teachers would give. Learners also said the system felt more supportive and fair, although some were worried about privacy and how their emotional data was used. This research helps us think about how future language tests could be more human-centered, taking into account not just what learners say, but how they feel while saying it.

Keywords

Artificial Intelligence, language assessment, human-centered AI, internet of things, oral proficiency, fairness in assessment

¹Muş Alparslan University, Turkey

Corresponding Author:

Sabahattin Yeşilçınar, Faculty of Education, Foreign Language Teaching Department, Muş Alparslan University, Mus 49250, Turkey.
Emails: sycinar@yahoo.com; sycinar@lparslan.edu.tr



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

Introduction

The integration of Artificial Intelligence (AI) and the Internet of Things (IoT) has rapidly advanced educational technologies, leading to increasingly sophisticated systems for instruction, feedback, and assessment. These developments have made the use of large language models (LLMs) in language teaching even more appealing. Thanks to LLMs, it has become possible to automate speaking assessments, create personalized learning pathways, and generate instantaneous scoring (Emirtekin, 2025). However, despite these advances, the affective dimensions of learning—emotions, stress, and engagement—which critically shape students' cognitive performance, are largely overlooked.

Regarding speaking skills in particular, it is well established that evaluating a language involves more than linguistic accuracy; it also requires managing anxiety, motivation, and interpersonal responsiveness. Contemporary validity frameworks (Bachman & Palmer, 2010; Weir, 2005a) place increasing emphasis on the characteristics of test takers and the consequences that assessments may produce (Xi, 2010). Yet in practice, test design often applies these principles only narrowly, prioritizing psychometric efficiency over learners' real-time emotional responses (Fulcher, 2015). One likely reason for this is the persistent mismatch between the elements that are theoretically endorsed and the realities of implementation.

Building on this gap, the present study proposes a responsive assessment framework that integrates affective computing, generative artificial intelligence, and IoT-based biofeedback to create emotionally attuned, fair, and valid evaluation environments sensitive to learners' real-time physiological and emotional states. The proposed responsive framework comprises three components: (a) multimodal emotion detection (HRV and facial expressions), (b) AI-supported scoring using GPT-4 and Whisper, and (c) a feedback system capable of adapting to the learner's emotional state during speaking tasks.

Literature Review

Generative AI in Speaking Assessment

AI, particularly LLMs such as GPT-4, has emerged as a promising force in language education. These systems are being integrated into language classrooms, supporting educators in creating original communicative tasks, generating natural dialogs, and providing adaptive feedback based on real-time student input (W. M. Lim et al., 2023). Moreover, generative AI enables the instantaneous and personalized creation of learning and assessment materials, which makes it highly applicable in both formative and summative assessment contexts. Moreover, these systems have offered dynamic and creative applications that go beyond traditional teaching and assessment formats

(Fui-Hoon Nah et al., 2023; Trust et al., 2023). This is because the technology can simulate dynamic oral and written tasks, adjust the level of difficulty according to the student's performance, and provide instant feedback that reflects not only linguistic accuracy but also discourse structure and communicative intent (Liang et al., 2023; Su & Yang, 2023).

Although these tools are generally discussed for their pedagogical benefits, a growing body of literature focuses explicitly on automated scoring. Studies in this area examine how well AI-produced scores align with those given by human raters and whether such systems can uphold standards of reliability, validity, and transparency. Recent research (Barrot, 2024; Bui & Barrot, 2025; Mizumoto & Eguchi, 2023) consistently emphasizes the need for scoring models to meet psychometric requirements and avoid systematic bias across learner groups. In this sense, generative AI is framed not as a task creation tool but as a scoring engine whose behavior must be carefully monitored and calibrated. Following this line of research, the present study explores whether adding affective information—such as HRV readings and facial emotion signals—can support fairness in automated oral scoring. Fairness is approached through two practical dimensions: procedural fairness, which relates to learners' understanding and control of data use; and interpretive fairness, which concerns whether emotion-aware scoring reduces construct-irrelevant penalties. Measurement invariance, although important in automated scoring, could not be investigated here due to sample size limitations.

Considering the questions raised about AI tools' alignment with human evaluative standards and their ability to account for contextual, emotional, and individual learner differences (Burstein et al., 2021; Floridi & Chiriatti, 2020), this study argues that generative AI should be viewed not as a replacement for human judgment in language assessment, but as a supportive tool embedded within an ethically guided learning ecosystem. In other words, it positions AI not merely as a tool, but as a component that participates in the learning process and helps create sensitive, fair, and pedagogically grounded assessment experiences. As language education continues to evolve in tandem with technological advances, the integration of generative AI into assessment should be critically and systematically conceptualized to ensure it supports learner agency, creativity, and meaningful communication.

IoT and Affective Computing in Educational Contexts

The integration of the IoT and *affective computing* into education has fundamentally changed the way learning environments are designed and implemented. In educational settings, IoT has evolved beyond mere

infrastructure improvement; it has become a structure that collects data on students' behavioral, cognitive, and physiological responses through smart classrooms equipped with sensors, wearable devices, and adaptive systems (Marquez et al., 2016). These systems offer dynamic and personalized learning experiences by adjusting the learning pace, content, and support provided according to students' individual learning profiles (Cheng & Liao, 2012). Smart watches, cameras, and microphones can now continuously monitor heart rate variability (HRV), facial expressions, and voice tone; all of these indicators are associated with cognitive load and stress (Picard, 2003). Such data sources enable affective computing systems that can interpret and respond to users' emotions; although promising in terms of personalized learning and participation monitoring, they are still underutilized in the assessment context. IoT refers to the ability of connected devices to communicate autonomously and exchange data, enabling real-time monitoring and analysis of user behavior in various fields, including education (Khan et al., 2021). IoT-based educational platforms have also supported language teaching by providing feedback on pronunciation and face/jaw movements using optical and sound sensors, while also enabling the data-driven adaptation of course content through Learning Management Systems (Ying, 2017).

In parallel, affective computing has also become a complementary element of IoT in educational technology. Affective computing refers to information processing that is related to emotion, derived from emotion, or deliberately influences emotion (Picard, 2003). In other words, it focuses on recognizing learners' emotional states and responding appropriately to those states. Affective computing technologies can detect emotional states such as confusion, anxiety, or interest by utilizing physiological signals, facial expressions, tone of voice, and similar multiple data sources, thereby enabling the adaptation of teaching strategies (Al-Azani & El-Alfy, 2025). The combined use of IoT and affective computing, particularly in language education and, more importantly, in the context of assessment, has the potential not only to improve learner performance but also to humanize the assessment process through real-time emotional awareness and interaction.

Validity, Fairness, and Human-Centered Assessment

Contemporary assessment theory has increasingly moved toward a broader understanding of validity. For example, Weir's (2005a) socio-cognitive framework argues that cognitive, contextual, and affective factors must be considered together in test design. Xi (2010) and Fulcher (2015) similarly emphasize that fairness—particularly for diverse learner profiles—is an ethical imperative.

However, traditional assessment methods have been criticized for their lack of standardization, issues with repeatability, and limited connection to real-world skills (Jacobs & Wallach, 2021), as well as for their inability to capture the contextual, nuanced, and human dimensions of language use. These concerns have prompted researchers to adopt evaluation frameworks that foreground human-centered values and contextual relevance.

A human-centered perspective views fairness not merely as consistent scoring, but as the design of systems capable of responding to different learner needs, emotional states, and interaction patterns. The integration of real-time affective feedback into assessments further aligns this study with current calls for more humane, inclusive, and equitable measurement practices. Transparency, interpretability, and learners' ability to contest or question automated outputs are foundational elements of ethical assessment. These principles resonate with longstanding concerns in the field of human-computer interaction (HCI), such as usability, emotional design, and participatory feedback loops (Olson & Kellogg, 2014). In educational contexts, user-friendly interfaces and emotionally aware feedback mechanisms not only enhance learner engagement but also contribute to perceptions of fairness and meaningfulness in assessment.

In AI-supported assessment, fairness extends beyond equal scoring. That is, it encompasses emotional accessibility, the transparency of feedback, adaptive interaction, and learners' rights to interpret AI-generated results (Liao & Xiao, 2023). Research by Langevin et al. (2021) on chatbots and writing assistants further shows that user autonomy and contextual appropriateness lie at the heart of ethical and pedagogically sound assessment systems. The literature on HCI and educational technology underscores the need to rethink validity and fairness in light of AI's expanding capabilities. Assessments today need to consider not only the learner's performance, but also how the learner interacts with the system, how they feel during that interaction, and what the resulting outcomes mean in real-world contexts. Human-centered assessment has therefore become not merely a methodological preference, but an ethical necessity. In this study, HCI refers to user-centered design, usability testing, participatory feedback cycles, contextual inquiry, and interpretive qualitative analysis. These methods, by prioritizing user autonomy, emotional experience, and interaction context, treat fairness not merely as a psychometric criterion but as a socio-technical phenomenon.

Theoretical Framework

This study draws upon an interdisciplinary theoretical framework to support the design and interpretation of the sentient assessment model:

Sociocognitive Validity Framework: A Guiding Lens for Emotion-Aware Assessment Design

Rather than focusing solely on language knowledge in isolation, Weir's sociocognitive framework, which considers the interaction between the candidate's cognitive abilities and the task context and how performance is assessed, offers a structured and evidence-based approach to examining the extent to which test tasks reflect real-life language use and human interaction. This makes it particularly suitable for studies (such as the current paper), where emotion, environment, and adaptive technologies are integral parts of the test experience.

Weir's framework consists of five interconnected components of test quality. *Cognitive validity* requires that tasks engage the same mental processes involved in authentic language use (Weir, 2005b). Accordingly, this study examined spontaneous speech production, planning under time pressure, and emotional self-regulation processes; word variety, average speech flow length, and complex sentence ratio were used as measures. These criteria were correlated with HRV (RMSSD) levels and anxiety indicators to examine how learners' emotional states interacted with their linguistic outputs. *Context validity* concerns the extent to which the assessment environment mirrors real communicative settings. Wearable sensors capture clues such as hesitation and stress by monitoring physiological responses that occur during speech. IoT-based systems provide continuous biofeedback, making the assessment environment more dynamic and interactive. *Scoring validity* relates to consistency and fairness. In this study, automated scores were compared with human scores to assess reliability, and it was expected that scores incorporating emotional information would show higher agreement with human evaluations compared to text-based scores alone (higher Pearson r and agreement coefficients). *Consequential validity* refers to the effects of an assessment on learners. Although the platform was experimental, factors such as perceived fairness, stress, and engagement were considered (Bridges, 2010). *Criterion-related validity* concerns how well the scores align with external measures. Although external correlations were not examined, emotion-informed scoring was expected to show a more consistent match with the exemplar performances used for calibration.

Unlike the more abstract, argument-based models of Bachman and Palmer (2010), Weir's framework does not rely on concepts such as chains of inference, which makes it more practical and transparent (G. Lim, 2013). The framework includes both a *priori* design evidence (task authenticity, prompt design, integration of emotion sensors) and a *posteriori* finding (learner feedback, score alignment). This distinction aligns well with the iterative and exploratory nature of the current study (Weir,

2005b). A key innovation in Weir's framework is his reframing of construct validity. Rather than defining language ability as a fixed trait within the individual, Weir (2005b) proposes that language performance should be seen as an interaction between cognitive ability and context. This interactionist perspective (see also Bridges, 2010) supports our use of affective and environmental data to better understand and support performance, especially under stress.

Human-Centered AI: Centering Learner Agency in Emotion-Aware Assessment

Shneiderman's (2022) model rethinks the role of artificial intelligence not as a replacement for human intelligence but as a tool that supports it. Instead of placing AI in the position of ultimate decision-maker, human-centered AI emphasizes shared responsibility, user adaptability, and ethical design principles. The fundamental principle is clear: AI should empower, not dominate. In educational applications, this means designing a language assessment platform that (i) respects learner emotions by integrating biofeedback (e.g., HRV, facial affect) to detect and respond to anxiety or stress in real time, (ii) supports transparency and control, allowing learners and educators to understand how AI scoring works and what influences outcomes, and (iii) maintains user trust through consistent, interpretable decisions and opportunities for learner feedback.

Traditional AI systems often prioritize performance metrics over user experience (Renz & Vladova, 2021; Xu, 2019). In contrast, human-centered design shifts the focus to how users feel, interact, and learn with AI. In this study, that includes:

- Using wearable IoT devices and generative AI (e.g., ChatGPT, Whisper) not just to evaluate speech, but to respond adaptively to learners' emotional states.
- Ensuring the platform is traceable and explainable, reducing the *black box* problem in AI (Arrieta et al., 2020; Guidotti et al., 2019).
- Encouraging collaboration between humans and machines, where learners are not passive test-takers, but co-constructors of their performance experience.

AI-based assessments are often based on technical accuracy, but this narrow focus ignores the emotional realities of language learners. A human-centered AI approach, based on Shneiderman's framework, reframes assessment goals: not simply scoring accuracy, but promoting learner well-being, dignity, and engagement throughout the assessment process. This perspective complements the

sociocognitive and dynamic assessment models in this study by offering a framework in which the AI does not merely assess language but collaborates with the learner in a fair, transparent, and emotion-sensitive manner. Following Weir (2005b), this study treated anxiety-induced disfluency as *construct-irrelevant variance* for general oral proficiency tests, which aim to measure linguistic ability, not emotion regulation. However, in constructs like *interactive competence under pressure*, emotion regulation was considered a relevant component. The system employed in this study sought to compensate for situational interference to approximate a “true score,” aligning with Kane’s (2013) emphasis on score interpretation rather than performance conditions.

Research Questions

This study adopts both a conceptual and empirical approach to exploring how generative AI, affective computing, and IoT devices can reshape the design and implementation of language assessment. Grounded in sociocognitive theory and human-centered AI principles, the study investigates how emotional and physiological responses influence speaking performance, how AI-generated scoring compares to human evaluation, and how learners perceive the fairness and emotional impact of such systems. It also critically examines the ethical implications of using real-time affective data in educational settings. To guide this investigation, the following research questions were formulated:

1. How do emotional and physiological indicators (e.g., HRV, facial expressions) related to learners’ oral language performance during AI-mediated speaking assessments?
2. To what extent can a sentient assessment system—integrating generative AI and affective computing—enhance scoring accuracy and alignment with human raters?
3. How do learners perceive the fairness, stress, and engagement of a sentient, emotion-responsive assessment platform in comparison to traditional oral evaluation methods?
4. What ethical considerations emerge from the use of real-time emotional and physiological data in language assessment?

Method

This study employed a mixed-methods exploratory design to evaluate the fairness, responsiveness, and learner perceptions of emotion-responsive language assessment platform. The platform integrated generative AI scoring (GPT-4), speech transcription (OpenAI

Whisper), affective computing, and smartwatches with HRV tracking to create a multidimensional evaluation system. The study aimed to examine how emotional signals influence speaking performance and how the integration of such data into AI scoring systems aligns with human judgment.

Participants

The study involved 24 undergraduate students (16 female, 8 male) enrolled in an oral communication course at a state school in Turkey. Participants were recruited through a course announcement and a follow-up email invitation. Inclusion criteria were (i) having a B1 to B2 CEFR proficiency level and (ii) reporting no cardiovascular conditions that could influence HRV measurements. To prevent potential novelty effects, individuals with prior experience using affective computing systems were excluded. No monetary compensation was offered. The study was approved by the Ethical Committee of the university of the author and was performed by the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments. Participation was entirely voluntary. All participants were informed about the study procedures, data types collected (audio, video, HRV, facial expression signals), data storage, and rights to withdraw. Written informed consent was obtained prior to participation. The consent form permitted opt-out, but all 24 participants ultimately consented to full sensing. To safeguard confidentiality, pseudonyms were used in all transcripts and reports. A subset of 18 students (11 female, 7 male) also volunteered for post-task semi-structured interviews to share their experiences and perceptions of the emotion-responsive assessment system.

A series of protective measures were implemented to minimize ethical risks in the study. First, only passive sensors were utilized and no physical intervention was applied. Participants were given complete freedom to pause or terminate the process at any time they wished. Facial expression data were processed locally on the device and permanently deleted after the analysis was completed. Similarly, HRV measurements were used only in summarized form; raw R–R intervals were not stored. By contrast, the benefits for participants were substantial: students had the opportunity to develop self-awareness of their own emotional responses, were introduced to innovative assessment technologies, and contributed to the development of fairer AI-based assessment practices.

Instruments and Technologies

The sentient assessment system comprised several tools and components:

- **Generative AI (GPT-4)** was used to generate oral prompts and automatically evaluate learner responses based on CEFR-aligned criteria.
- **OpenAI Whisper**, a speech-to-text tool, transcribed participants' oral responses in real time. Its high accuracy enabled seamless integration into the AI scoring pipeline.
- **Facial emotion recognition:** A Logitech C920 webcam (1080p/30 fps) recorded facial expressions under standardized indoor lighting (350–450 lux). Expressions were categorized based on Ekman's (1992) universal emotion taxonomy. Although Ekman's model has limitations, it remains widely used in real-time expression analysis. OpenFace 2.0 (Baltrušaitis et al., 2016) was used for frame-by-frame Action Unit analysis, achieving 82% to 87% accuracy on common benchmark datasets (e.g., BP4D, DISFA). Only categorical emotion labels (e.g., Neutral, Anxious, Surprised) were stored.
- **Smartwatches with HRV tracking:** Devices recorded HRV as an indicator of emotional arousal. RMSSD, the standard short-term HRV metric, was computed using 60-s rolling windows of PPG-derived R–R intervals. Lower HRV values indicated heightened stress or anxiety.
- **Human raters ($n = 3$):** Three MA-level English teachers with 5 to 8 years of rating experience evaluated all recordings. A 3-hr norming session using five pilot samples yielded inter-rater reliability of $\alpha > .80$. Raters independently scored 72 performances (24 participants $\times$ 3 tasks) using a CEFR-based analytic rubric. Identities and emotional data were concealed. Fifteen percent of samples were double-scored to determine intra-rater reliability ($r = .85$).

Procedure

The assessment sessions were conducted in a digitally equipped language lab. The procedure began with a piloting phase conducted to ensure the feasibility and reliability of the integrated assessment system. The piloting phase involved four volunteer students (three female, one male; CEFR B1–B2; mean age 20.5) from the same cohort who did not participate in the main study. Pilot data were excluded from all analyses. During piloting, a small group of participants completed the speaking tasks under the full setup to test the synchronization of biometric data collection, speech transcription, emotion recognition, and real-time feedback. Adjustments were made based on pilot results to optimize system responsiveness and participant comfort. For the main sessions, participants were fitted with smartwatches to monitor

HRV, and a baseline resting HRV was recorded for 3 min. Participants then received speaking prompts generated by GPT-4, displayed on a screen, and responded orally. Their speech was transcribed in real time using OpenAI Whisper technology, enabling immediate linguistic analysis by GPT-4 to generate initial proficiency scores. Transcribing participants' speech was essential to produce an accurate, detailed linguistic assessment and to feed the AI scoring system with precise input. Concurrently, a webcam captured facial expressions to analyze emotional states, and HRV data provided physiological indicators of stress or emotional arousal. The AI system then integrated linguistic, emotional, and physiological data to produce an emotion-informed score that accounted for participants' stress levels during the task. Each speech sample underwent two separate automated scoring processes. In the first stage, GPT-4 produced an evaluation based solely on the speech transcript according to CEFR criteria. In the second stage, the same transcript was presented to the model together with emotional metadata derived from HRV measurements and facial expression analysis, generating an emotion-sensitive score. Participants received real-time feedback in the form of a color indicator, optional breathing guidance, and a pause button. The real-time feedback followed a simple three-color logic: green (stable HRV and neutral affect), yellow (mild arousal), and red (sustained low HRV or anxious affect). When red was detected, the interface displayed an optional breathing prompt. These features were intentionally minimal to reduce cognitive load; they did not require user interaction and were positioned in peripheral vision. As such, emotional feedback was designed to be supportive rather than intrusive, minimizing the risk of additional anxiety. While human raters evaluated the videos directly, GPT-4 partially compensated for fluency losses by augmenting transcripts with Whisper timestamps and emotional metadata. Because transcripts alone cannot represent pausing behavior or micro-disfluencies, Whisper's timestamps were included to preserve pause length and frequency. These timestamps allowed GPT-4 to approximate fluency by interpreting temporal gaps, although human raters still retained access to richer prosodic features. This approach ensured that GPT-4's fluency judgments were grounded in measurable timing-based signals rather than inferred prosody. Real-time feedback was presented visually to participants, offering adaptive prompts to support emotional regulation, such as breathing suggestions when elevated stress was detected. Although emotional signals were continuously collected, the distinction between pre- and post-emotional input was made analytically rather than experimentally. Transcript-only scoring occurred immediately after each spoken turn, while affect-informed scoring was generated in a second pass using

the same transcript combined with the emotional metadata. Consequently, “pre-emotional input” corresponds to transcript-only scores, whereas “post-emotional input” corresponds to the emotion-informed re-scoring step. GPT-4 distinguished the two scoring stages through separate prompt calls containing either (a) no emotional metadata tokens or (b) the metadata tokens added to the prompt. No biological or emotional data were streamed into the first call, ensuring the model could not access affective information during the transcript-only scoring stage. After completing the tasks, some participants took part in semi-structured interviews to share their perspectives on the system’s fairness, emotional comfort, transparency, and control.

Data Collection and Analysis

Quantitative data were gathered through biometric tools (e.g., HRV and facial expression tracking software) and performance scores obtained from both human raters and AI-based systems. AI-human score comparisons were analyzed using Pearson’s r to assess correlation, while Krippendorff’s alpha measured inter-rater reliability among human raters. Pre- and post-emotional input AI scores were compared to assess improvement in interpretive fairness; however, this comparison is an analytical distinction, not an experimental manipulation, and the timing procedure was not systematically recorded.

For the qualitative component, semi-structured interviews were conducted with participants to explore their perceptions and experiences regarding sentient features in online assessment. The interview data were transcribed verbatim and coded using MAXQDA. An inductive thematic analysis approach was used to analyze the transcripts, following Braun and Clarke’s (2006) six-step procedure. This analysis identified four prominent themes: emotional support, fairness perceptions, surveillance concerns, and learner agency. These themes reflect participants’ subjective interpretations of how sentient assessment features shaped their engagement, motivation, and perceived autonomy. Triangulation of the qualitative themes with biometric and scoring data allowed for a comprehensive understanding of the interplay between affective, perceptual, and performance dimensions in technology-mediated assessment contexts.

Findings

Emotional and Physiological Signals in Relation to Oral Performance

When the findings were examined, it was observed that the emotional and physiological states of the learners

affected their speech performance during the AI-mediated task. In other words, participants with lower HRV values and negative affective indicators (e.g., jaw tension, narrowed eyes, furrowed brows) paused more frequently during speech and continued to speak with hesitant and relatively shorter sentence structures. In addition, it was observed that these students had a limited vocabulary and avoided complex syntactic structures.

For example, a participant whose stress levels increased at the beginning of the task showed a significant decrease in HRV values, and the facial expression software consistently detected anxiety. This student’s speech was frequently interrupted by pauses and repetitive starters such as “*I think... I think... maybe it’s like...*” She later reflected, “*I could feel my heart beating too fast, and my brain just went blank. I forgot what I wanted to say, even though I practiced it yesterday.*” (Daisy)

Another participant was found to exhibit more balanced physiological responses, consequently producing more fluent and structurally longer expressions. This participant maintained a calm and neutral facial expression and demonstrated a high level of control in the use of conditional structures and discourse modifiers. He described their experience as follows: “*I felt more confident because it didn’t feel like a test. It was just me and the system, and it didn’t judge me with a face.*” (Michael)

Interestingly, several learners displayed emotional recovery during the task, with stress indicators decreasing midway through the session. These learners often responded well to adaptive system prompts, such as the option to pause or slow down. As one learner described, “*When the system showed that my stress was high and told me to take a short breath, it actually helped. I started to think more clearly and said what I meant.*” (Rose)

These patterns suggest that short-term affective shifts—especially reductions in anxiety—can meaningfully influence language performance, even within the span of a few minutes. As shown in Table 1, a clear pattern emerged across participants: high physiological stress (e.g., low HRV, anxious expressions) was consistently associated with disfluency, lexical avoidance, and shorter utterances, while more emotionally stable students produced extended, coherent speech with greater syntactic and lexical complexity.

Overall, Table 1 reinforces that emotional and physiological data can meaningfully influence how speaking performance is manifested and interpreted. Learners who experienced real-time stress reduction demonstrated improved fluency and coherence mid-task, especially when supported by AI prompts. This suggests that emotional regulation is not only correlated with but may be predictive of language output quality during high-stakes oral assessments

Table 1. Emotional and Physiological Indicators Versus Performance Outcomes.

Emotional/physiological state	Observed linguistic performance	AI response (no emotion input)	AI response (with emotion input)
Low HRV, anxious expression	Disfluency, lexical avoidance, short utterances	Penalized heavily	Fluency adjusted
Stable HRV, neutral affect	Coherent, extended turns, varied lexis and syntax	Accurate scoring	Accurate scoring
Emotional recovery mid-task	Improved coherence and self-correction	No recognition	Recognized as an adaptation

Table 2. Comparison of AI Versus Human Ratings Across Scoring Dimensions.

Scoring dimension	Human inter-rater reliability (α)	AI-human correlation (r)	AI-human correlation (emotion input)
Fluency	0.81	.78	.83
Coherence	0.84	.76	.82
Lexical range	0.80	.75	.77
Pronunciation	0.79	.74	.75

AI Scoring and Alignment with Human Ratings

The comparison between GPT-4-generated scores and human rater evaluations ($n = 3$) demonstrated a high level of agreement across key speaking dimensions, including fluency, coherence, lexical range, and pronunciation. The Pearson correlation between AI scores and the average of human scores was $r = .78$ ($p < .01$), indicating a strong overall alignment. Moreover, inter-rater reliability among human evaluators was robust, with a Krippendorff's α of .81, confirming consistency across human assessments. However, a closer analysis revealed important discrepancies in how disfluencies—particularly those related to anxiety—were interpreted. In one notable case, a learner who exhibited vocal tremors and hesitation due to visible stress received a low fluency score from the initial AI model. Human raters, by contrast, acknowledged the emotional difficulty but evaluated the performance more holistically, emphasizing the learner's ability to maintain topic relevance and semantic clarity despite occasional pauses.

When emotional and physiological data were incorporated into GPT-4's scoring prompts, the model's assessments became more sensitive to anxiety-related disfluency. Mixed-effects modeling showed that adding affective features significantly improved model fit ($\chi^2(2) = 10.9$, $p = .004$). Marginal R^2 increased from .42 to .48, yielding $\Delta R^2 \approx .06$. This emotion-informed version of the AI achieved an even stronger correlation with human ratings, particularly for fluency and coherence, reaching $r = .83$ (see Table 2). Using Fisher's z -transformation, the 95% confidence interval for the transcript-only AI-human correlation ($r = .78$, $N = 24$) was [.56,

.89]. The 95% CI for the emotion-informed correlation ($r = .83$) was [.65, .92]. These intervals confirm that the improvement in alignment was statistically meaningful. In these cases, the model adjusted its interpretation of disfluency—not as a deficit in language ability, but as a reflection of cognitive or emotional load.

The adjusted scoring mechanism was developed by embedding real-time emotional metadata into GPT-4's prompt structure. Inputs such as [HRV: Low], [Emotion: Anxious], and [Stress Recovery: Detected] were included alongside transcription data. This allowed GPT-4 to distinguish between disfluencies caused by language limitations and those arising from temporary affective disruption. As a result, the model could more accurately assess speaker intent and coherence under emotionally challenging conditions.

Student reflections reinforced these findings. Adam shared, “*I think the AI was fairer when it understood I was nervous. At first, I got a low score. But after the update, it matched what the teacher said.*” Another student, Adar, expressed a similar sentiment: “*It felt like the system listened to my effort, not just my words. That made a big difference for me.*” These observations suggest that emotion-informed scoring systems offer a more nuanced and human-like assessment of performance—particularly valuable in high-stakes oral assessments where anxiety can impact learner output. Integrating affective data into scoring systems enhances fairness by accounting for the cognitive and emotional realities of speaking under pressure.

Learner Perceptions of Fairness, Support, and Emotional Impact

Participants generally responded positively to the sentiment features of the platform, highlighting the adaptive feedback and visible stress monitoring as key elements that made the experience feel less judgmental. Many described the system as non-threatening, encouraging, or more forgiving than a traditional oral exam. Suzan reflected, “*It was different because the system could see I was nervous and gave me time. In a normal exam, the teacher just waits, and I panic more.*” Another participant expressed appreciation for being able to see their emotional status in real-time, “*The color indicator made me*

realize I was too tense. I started breathing slower after that, and it helped me say what I meant.” (Deniz).

However, not all reactions were uniformly positive. A few learners expressed concern about being emotionally monitored, especially by a machine. Senem noted, *“I didn’t like the idea that my face was being watched and analyzed. I felt exposed, like I had no control over how the system saw me.”* Another participant shared skepticism, *“Can AI really understand how I feel? What if it makes a wrong guess and that affects my score?”* (Mine). Still, the majority valued the consistency of AI-generated scores and the absence of perceived human bias. Several noted that the system *felt fairer* than past oral exams, where they felt disadvantaged by unfamiliar assessors or subjective interpretations of their accent.

Ethical Considerations: Trust, Consent, and Learner Agency

Ethical concerns emerged clearly in learner interviews, especially regarding data transparency, emotional consent, and future use of their physiological information. While most participants agreed that emotional responsiveness improved the experience, many emphasized the need for clear explanations of what data was being collected, how it was used, and whether it influenced scores. Sena student emphasized, *“If the system uses my feelings to score me, I should know exactly how. It’s my data, and I should understand the decision.”* Kaya, who initially felt positive about the experience, later reflected, *“It’s fine for practice, but I wouldn’t want this in a real exam. I don’t want my heart rate stored somewhere I can’t control.”*

Some learners expressed discomfort with what they called the *invisibility* of the data collection process, stating that the sensors felt *silent but powerful* (e.g., Asmin, Halime, Çınar, and Nedim). Suggestions from participants included the ability to review and delete their emotion data after testing, being able to *opt out* of emotional analysis in high-stakes settings, including human moderation in final scoring decisions. Despite these concerns, most learners supported the platform’s use in low-stakes or formative contexts, particularly when framed as a self-awareness tool rather than an evaluator. A few even requested continued access to emotional feedback for personal tracking.

Discussion

This section interprets the study’s findings through the lenses of Weir’s sociocognitive validity framework and Shneiderman’s model of human-centered AI. Each subsection parallels the structure of the Findings section to ensure conceptual clarity and theoretical alignment.

Emotional and Physiological Signals in Relation to Oral Performance

The integration of IoT and affective computing technologies—such as HRV monitors and facial emotion recognition—allowed this study to uncover strong links between learners’ emotional states and their oral language performance. These findings affirm Weir’s (2005a) notion of context and cognitive validity, where performance should be interpreted in light of both the linguistic task and the test-taker’s cognitive-emotional state. As supported by Picard (2003), physiological responses such as increased heart rate and negative affect correlate with cognitive overload and reduced output fluency.

Learners with fluctuating or elevated stress exhibited more fragmented speech and lexical avoidance, while those with stable physiological-emotional profiles demonstrated greater fluency and syntactic control. These outcomes support claims by Chang and Chen (2023) and Al-Azani and El-Alfy (2025) that emotionally aware systems can positively mediate cognitive engagement. Additionally, the platform’s adaptive feedback—such as stress-regulated pacing—exemplifies the principles of the Cognitive-Affective Theory of Learning with Media (CATLM), which emphasizes the interaction between affective design elements and learner motivation and retention (Chang & Chen, 2023), by creating emotionally responsive learning experiences. Crucially, the detection of emotional shifts and recovery during task execution underlines the value of using real-time affective signals not as passive indicators but as interactive scaffolding cues, reinforcing the dynamic nature of assessment.

AI Scoring and Alignment with Human Ratings

The study demonstrated that generative AI, particularly GPT-4, can approximate human scoring with relatively high correlation, especially when enhanced by emotional context. Initially, the AI system aligned with prior concerns in the literature (Burststein et al., 2021; Floridi & Chiriatti, 2020) by misinterpreting anxiety-induced hesitations as cognitive deficiencies. However, when bio-affective inputs were integrated, the system recalibrated its interpretation of disfluencies, treating them not as performance gaps but as emotionally-driven phenomena. This resulted in more equitable scoring, consistent with sociocognitive and scoring validity (Weir, 2005a). Within Weir’s framework, the incorporation of affective data strengthens scoring validity by reducing construct-irrelevant variance caused by anxiety-induced disfluency. It also supports context validity, as performance is interpreted in relation to the test-taker’s real-time cognitive-emotional state—a core component of Weir’s model. Thus, affective integration reinforces the interpretive

argument connecting performance, scoring, and the test context.

This shift echoes recent calls to position generative AI not as a scoring authority but as a supportive tool in a human-centered assessment ecosystem (Galaczi, 2023; Shneiderman, 2022). GPT-4's adaptive capabilities—particularly when informed by multimodal data—enabled it to interpret not just what learners said, but how and under what emotional constraints. This aligns with Liang et al. (2023), who argue that discourse structure and communicative intent should guide AI evaluation. Still, cultural authenticity and idiomatic nuance—longstanding concerns in AI-generated feedback (Craig, 2022)—remain limitations. These findings suggest that while generative AI offers speed and objectivity, it should be situated within transparent, emotionally aware, and human-interpretable frameworks.

Unlike conventional AI scoring tools, this study introduces an affect-augmented scoring process. While traditional systems excel in consistency, they do not account for moment-to-moment emotional states that can depress fluency or lexical retrieval. The present system, therefore, extends existing models by incorporating real-time affective context into the scoring pipeline.

Learner Perceptions of Fairness, Support, and Emotional Impact

Learners' reflections highlighted a shift in how fairness is understood in AI-mediated assessment. Many viewed the sentient system as more supportive and transparent than traditional oral exams. These responses support Xi (2010) and Fulcher (2015), who argue that fairness involves more than psychometric balance—it involves learners' subjective sense of justice, agency, and engagement.

Participants noted that adaptive emotional feedback (e.g., pause prompts or calm-tone suggestions) reduced performance anxiety and enhanced focus. These outcomes support CATLM and the use of affective computing in promoting emotional accessibility in learning environments. However, several learners expressed unease about being emotionally monitored. This concern reflects wider anxieties in affective AI and surveillance ethics (Mehrabian et al., 2022; Shneiderman, 2020), and underscores the need for opt-in consent, transparency of data use, and contestability of AI-generated outcomes (Liao & Xiao, 2023). The findings suggest that emotional atonement should be balanced with data ethics and user autonomy—a cornerstone of human-computer interaction (Olson & Kellogg, 2014).

Ethical Considerations: Trust, Consent, and Learner Agency

The deployment of AI systems that collect and interpret emotional and physiological data introduces significant ethical considerations. While participants appreciated the responsiveness of the platform, several questioned how their data were used and stored. These responses reinforce the importance of human-centered AI principles (Shneiderman, 2022), particularly regarding explainability, user control, and shared accountability.

Unlike earlier models that emphasize predictive accuracy (Xu, 2019), human-centered AI prioritizes user experience, emotional impact, and interpretability. In this study, that meant designing a system that allowed learners to view emotional feedback in real time and understand how it influenced scoring. Nevertheless, some participants felt that the affective layer introduced pressure rather than relief, suggesting that emotional input should be an adjustable and learner-controlled feature, not a fixed analytic layer. Furthermore, in alignment with Jacobs and Wallach (2021) and Langevin et al. (2021), this study emphasizes that assessment systems must go beyond collecting affective data—they must return interpretive power to the learner. Ethical assessment must offer learners not just feedback, but the means to understand, contextualize, and challenge it. Only through such practices can AI-mediated systems evolve into trustworthy, human-aligned assessment partners.

To synthesize the findings and their theoretical grounding, Figure 1 presents the Sentient Assessment Framework. This model positions emotional signals, generative AI, and IoT technologies as foundational, interacting inputs that enable the emergence of adaptivity, interpretability, and fairness in language assessment. These outcomes reflect not only psychometric concerns but also learners' emotional experience, agency, and trust in the system.

This framework illustrates how the integration of emotional signals, generative AI, and IoT-based data collection operates as a composite input system for sentient assessment. These elements work synergistically to support key outcomes—adaptivity, interpretability, and fairness—in real-time language evaluation. The upward directional flow highlights the formative relationship between sensor-based emotional awareness and human-centered assessment goals. This model builds on Weir's sociocognitive validity and Shneiderman's principles of human-centered AI to propose a future-facing design structure for emotion-responsive, ethically attuned digital assessment systems.

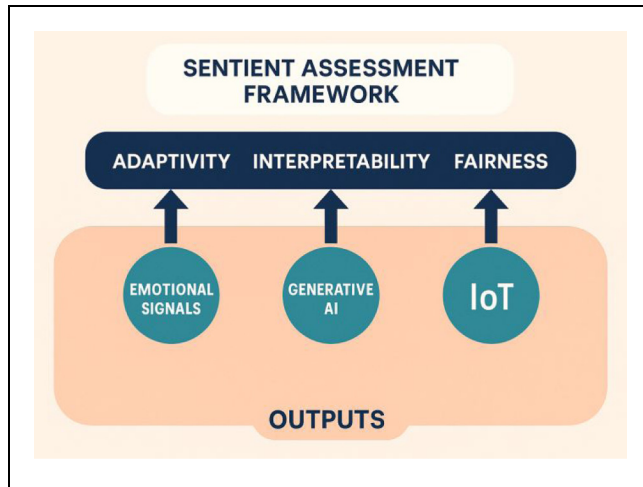


Figure 1. The sentient assessment framework: inputs and outcomes.

Conclusion

This study introduced a sentient, AI-mediated language assessment model that integrates affective computing and IoT biofeedback to support emotionally responsive and human-centered evaluation. Findings from the pilot study affirm that learners' emotional states directly influence oral performance and that AI scoring becomes more valid and empathetic when contextualized by real-time affective data. Learners responded positively to the adaptive features, though some raised important ethical concerns about emotional privacy and algorithmic opacity.

Theoretically, the study contributes to expanding sociocognitive and dynamic models of assessment by positioning emotion as a core construct, not an external confound. Practically, it demonstrates how generative AI systems can shift from mechanistic scoring toward interpretive, learner-responsive platforms that foster trust, reduce anxiety, and promote fairness—especially for learners who struggle with performance pressure.

At a time when AI systems are being rapidly embedded into educational infrastructures, this research provides a timely and critical model for how language assessment can be both technologically advanced and ethically attuned. It argues for a vision of AI not as a dispassionate evaluator but as a supportive tool in co-constructing meaning, well-being, and opportunity for learners.

Future Directions

While promising, this study is exploratory and limited by sample size and context (intermediate learners in a pilot environment). Future research should explore longitudinal effects of emotion-responsive assessment systems,

their application in different cultural and linguistic settings, and the boundaries of learner comfort with biometric feedback. Although the study used a within-performance comparison, a fully counterbalanced or independent control condition would enable stronger causal conclusions. Future work should incorporate such designs. In addition, integrating teacher-facing dashboards, emotion trace logs, and learner feedback loops could further enhance transparency and collaborative use. Finally, as AI's role in education expands, ongoing attention should be paid to the governance, interpretability, and democratization of such systems. Sentient assessment must remain rooted in pedagogy—not surveillance—and should be built not only for accuracy, but for empathy, dignity, and learner growth.

ORCID iD

Sabahattin Yeşilginar  <https://orcid.org/0000-0001-6457-0211>

Ethical Considerations

The study was approved by the Ethical Committee of the Muş Alparslan University and was performed by the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Supplemental Material

Supplemental material for this article is available online.

References

- Al-Azani, S., & El-Alfy, E. S. M. (2025). A review and critical analysis of multimodal datasets for emotional AI. *Artificial Intelligence Review*, 58(10), 334. <https://doi.org/10.1007/s10462-025-11271-1>

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. OUP Oxford.
- Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). *Openface: An open source facial behavior analysis toolkit* [Conference session]. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). <https://doi.org/10.1109/FG.2018.00019>
- Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, 29(6), 7155–7179. <https://doi.org/10.1007/s10639-023-12083-y>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Bridges, G. (2010). Demonstrating cognitive validity of IELTS academic writing task. *University of Cambridge ESOL Examinations Research Notes*, 42, 24–33.
- Bui, N. M., & Barrot, J. (2025). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. *Innovation in Language Learning and Teaching*. Advance online publication. <https://doi.org/10.1080/17501229.2025.2521003>
- Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2021). A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test. *Duolingo Research Reports (DRR-22-01)*. <https://duolingo-papers.s3.amazonaws.com/other/det-assessment-ecosystem.pdf>
- Chang, C. C., & Chen, T. C. (2023). Effects of digitization/mobilization and emotional design of textbooks on emotion, various cognitive load and cognitive performance. *Education and Information Technologies*, 28(11), 14827–14855. <https://doi.org/10.1007/s10639-023-11779-5>
- Cheng, H. C., & Liao, W. W. (2012). *Establishing an lifelong learning environment using IOT and learning analytics* [Conference session]. 2012 14th International Conference on Advanced Communication Technology (ICACT).
- Craig, A. (2022). Interrogating racialized “Cultural Authenticity” discourses among language-learner tourists in Australia. *Tourism Management*, 90, 104468. <https://doi.org/10.1016/j.tourman.2021.104468>
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3–4), 169–200. <https://doi.org/10.1080/02699939208411068>
- Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. *Applied Sciences*, 15(10), 5683. <https://doi.org/10.3390/app15105683>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304. <https://doi.org/10.1080/15228053.2023.2233814>
- Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.
- Galaczi, E. (2023). *English language education in the era of generative AI: Our perspective*. Cambridge University Press & Assessment. <https://www.cambridgeenglish.org/Images/685411-english-language-education-in-the-era-of-generative-ai-our-perspective.pdf>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining Black Box Models. *ACM Computing Surveys*, 51(5), 1–42. <https://doi.org/10.1145/3236009>
- Jacobs, A. Z., & Wallach, H. (2021). *Measurement and fairness* [Conference session]. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. <https://doi.org/10.1145/3442188.3445901>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Khan, R., Ali, I., Jan, M. A., Zakarya, M., Khan, M. A., Alshamrani, S. S., & Guizani, M. (2021). A hybrid approach for seamless and interoperable communication in the internet of things. *IEEE Network*, 35(6), 202–208. <https://doi.org/10.1109/MNET.011.2000787>
- Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021). *Heuristic evaluation of conversational agents* [Conference session]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3411764.3445312>
- Liang, J. C., Hwang, G. J., Chen, M. R. A., & Darmawansah, D. (2023). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*, 31(7), 4270–4296. <https://doi.org/10.1080/10494820.2021.1958348>
- Liao, Q. V., & Xiao, Z. (2023). Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*. <https://arxiv.org/abs/2306.03100>
- Lim, G. (2013). Components of an elaborated approach to test validation. *Cambridge English: Research notes*, 51, 11–14. <https://www.cambridgeenglish.org/Images/130828-research-notes-51-document.pdf#page=13>
- Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *International Journal of Management in Education*, 21(2), 100790. <https://doi.org/10.1016/j.ijme.2023.100790>
- Marquez, J., Villanueva, J., Solarte, Z., & Garcia, A. (2016). IoT in education: Integration of objects with virtual academic communities. In Á. Rocha, A. Correia, H. Adeli, L. Reis, & M. Mendonça Teixeira (Eds.), *New advances in information systems and technologies* (pp. 201–212). Springer. https://doi.org/10.1007/978-3-319-31232-3_19
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>

- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Olson, J. S., & Kellogg, W. A. (Eds.) (2014). *Ways of knowing in HCI* (Vol. 2, p. 74). Springer.
- Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies*, 59(1–2), 55–64. [https://doi.org/10.1016/s1071-5819\(03\)00052-1](https://doi.org/10.1016/s1071-5819(03)00052-1)
- Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered artificial intelligence in Educational Technologies. *Technology Innovation Management Review*, 11(5), 5–16. <https://doi.org/10.22215/timreview/1438>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.
- Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, 6(3), 355–366. <https://doi.org/10.1177/20965311231168423>
- Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, 23(1), 1–23. <https://www.learntechlib.org/primary/p/222408/>
- Weir, C. J. (2005a). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281–300. <https://doi.org/10.1191/0265532205lt309oa>
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147–170. <https://doi.org/10.1177/0265532209349465>
- Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, 26(4), 42–46. <https://doi.org/10.1145/3328485>
- Ying, Y. (2017). Application of flipped classroom teaching mode based on MOOC in modern educational technology teaching. *Journal of Computational and Theoretical Nanoscience*, 14(2), 1075–1078. <https://doi.org/10.1166/jctn.2017.6405>