Computers and Education: Artificial Intelligence 9 (2025) 100436

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

GLAT: The generative AI literacy assessment test

Yueqiao Jin, Roberto Martinez-Maldonado, Dragan Gašević, Lixiang Yan *

Monash University, 25 Exhibition Walk, Clayton, 3168, Victoria, Australia

ARTICLE INFO

Keywords:
Generative AI
AI literacy
Item response theory
Classical test theory
Higher education
Validity and reliability
Learning performance

ABSTRACT

The rapid integration of generative artificial intelligence (GenAI) technology into education requires precise measurement of GenAI literacy to ensure that learners and educators possess the skills to engage with and critically evaluate this transformative technology effectively. Existing instruments often rely on self-reports, which may be biased. In this study, we present the GenAI Literacy Assessment Test (GLAT), a 20-item multiple-choice instrument developed following established procedures in psychological and educational measurement. Structural validity and reliability were confirmed with responses from 355 higher education students using classical test theory and item response theory, resulting in a reliable 2-parameter logistic (2PL) model (Cronbach's alpha = 0.80; omega total = 0.81) with a robust factor structure (RMSEA = 0.03; CFI = 0.97). Critically, GLAT scores were found to be significant predictors of learners' performance in GenAI-supported tasks, outperforming self-reported measures such as perceived ChatGPT proficiency and demonstrating external validity. These results suggest that GLAT offers a reliable and valid method for assessing GenAI literacy, with the potential to inform educational practices and policy decisions that aim to enhance learners' and educators' GenAI literacy, ultimately equipping them to navigate an AI-enhanced future.

1. Introduction

Generative artificial intelligence (GenAI) has rapidly emerged as a transformative force in higher education, challenging traditional pedagogical frameworks while simultaneously presenting novel opportunities for teaching, learning, and assessment. Tools like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude hold the potential to transform how personalised tutoring service can be delivered, how instructional materials can be generated, how lectures can be transcribed for accessibility, and how creativity can be nurtured through multimedia content generation (Yan et al., 2024a; Khosravi et al., 2023). However, the integration of these technologies is accompanied by complex challenges, including ethical considerations, risks of misinformation from model "hallucinations," and concerns regarding academic integrity (Ji et al., 2023; McDonald et al., 2024). Such complexities necessitate a deeper focus on fostering AI literacy, particularly GenAI literacy, among both educators and learners to fully harness GenAI's benefits and mitigate its associated risks (Ng et al., 2021b; Zhao et al., 2024).

AI literacy refers to the set of competencies that enable individuals to effectively interact with AI technologies, encompassing understanding fundamental AI concepts, engaging in critical evaluation, and using AI tools ethically in diverse contexts (Long & Magerko, 2020; Ng et al., 2021b). Within this broad framework, GenAI literacy emerges as a specialised subset, focusing on skills required to engage with GenAI systems that can autonomously produce text, visuals, or other forms of media (Yan et al., 2024a; Annapureddy et al., 2024). Developing GenAI literacy involves more than just foundational knowledge; it requires proficiency in crafting prompts, interpreting AI-generated outputs, and understanding the socio-ethical implications of using such tools (Zhao et al., 2024; Bozkurt, 2024a). As GenAI becomes increasingly embedded in educational systems, it is imperative for learners and educators to acquire these competencies to effectively leverage the technology while minimising potential pitfalls such as biases or inaccuracies (Lyu et al., 2024; Chiu, 2024).

Numerous instruments have been developed to assess AI literacy, reflecting the diversity of competencies that individuals need to navigate AI technologies. Conventional AI literacy assessments often rely on self-reported surveys, which are effective in capturing perceived knowledge but may lack the reliability needed to accurately measure actual competencies, especially given the tendency for individuals to overestimate their understanding (Lintner, 2024; Laupichler et al., 2023b). Most existing instruments address general AI literacy, focusing on technical knowledge, awareness, and ethical considerations, but fail to adequately capture the unique skills required for GenAI (Koch et al., 2024;

* Corresponding author.
E-mail address: lixiang.yan@monash.edu (L. Yan).

https://doi.org/10.1016/j.caeai.2025.100436
Received 22 November 2024; Received in revised form 11 May 2025; Accepted 4 June 2025
Available online 9 June 2025

Zhao et al., 2024). There is a growing demand for more nuanced and context-specific instruments to evaluate GenAI literacy, particularly as generative tools become integral to both physical and digital learning environments (Koch et al., 2024; Zhao et al., 2024; Yan et al., 2024b).

Current AI literacy assessments can be broadly categorised into two types: self-reported and performance-based measures. Self-reported instruments, while commonly used, provide insights into individuals' perceived abilities but may introduce biases that obscure a more reliable measure of literacy levels (Ng et al., 2021b; Lintner, 2024). In contrast, performance-based assessments evaluate actual competencies through direct engagement, offering a more reliable measure of skills. This distinction is especially pertinent for GenAI literacy, where there is often a gap between learners' perceived understanding and their real ability to effectively utilise generative tools (Lyu et al., 2024). GenAI technologies necessitate iterative, context-specific interactions that require both sophisticated prompting skills and the ability to critically assess AI outputs, areas where self-reports may fall short (Chiu, 2024; Bozkurt, 2024a). Therefore, developing performance-based instruments is essential to provide a reliable assessment of individuals' abilities to engage with these advanced technologies in educational settings. However, to the best of our knowledge, there are still limited performance-based tools for measuring students' GenAI literacy in higher education, particularly those that have been rigorously developed and validated according to established psychological and educational measurement standards (Thorndike et al., 1991; American Educational Research Association et al., 2014).

The current study contributes to the field of AI in education and AI literacy by introducing the GenAI Literacy Assessment Test (GLAT), a performance-based instrument specifically designed to evaluate GenAI literacy within higher education contexts. The GLAT aims to fill a critical gap in existing assessment tools by providing a more reliable, comprehensive evaluation of the key competencies required to interact with GenAI tools. Unlike existing assessments that focus predominantly on general AI skills, GLAT targets the unique skills necessary for effective engagement with generative technologies, including technical proficiency, ethical awareness, and the capacity for critical evaluation of GenAI-generated outputs. This instrument is grounded in rigorous methodologies from psychological and educational measurement (Thorndike et al., 1991; American Educational Research Association et al., 2014), ensuring both validity and reliability. By focusing on performance-based metrics, GLAT provides educators and researchers with a reliable tool to assess how well students and educators understand and can leverage GenAI technologies, ultimately informing targeted interventions that can enhance these competencies.

2. Background

2.1. AI and GenAI literacy

The rapid advancements in AI technologies have accentuated the importance of AI literacy, especially in educational research. Researchers have endeavoured to define and conceptualise "AI literacy," with the definition by Long and Magerko (2020, p.2) being frequently cited: "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace." Building upon this definition, other researchers have explored essential aspects of AI literacy. For instance, Kandlhofer et al. (2016) and Burgsteiner et al. (2016) concentrated on the comprehension of fundamental AI concepts present in various products and services. Meanwhile, Wang et al. (2023) emphasised critical evaluation, practical application, and ethical responsibilities. Additionally, Ng et al. (2021b), Ng et al. (2021a), and Almatrafi et al. (2024) have refined this framework by stressing competencies such as recognition, application, evaluation, creation, and ethical navigation.

The rise of GenAI technologies, like ChatGPT, necessitates a re-evaluation of AI literacy within the specific context of generative technologies. GenAI's capability to generate substantial content from minimal input alters the landscape, prompting a need for a revised understanding of AI literacy in this context (Zhao et al., 2024). Although there has been increasing academic interest in GenAI, a comprehensive definition of GenAI literacy remains elusive (Annapureddy et al., 2024). Many current AI literacy frameworks are too general and do not address the specific competencies required by GenAI, which differ significantly from those of predictive models. Specifically, GenAI literacy calls for an integrative approach that combines theoretical knowledge, practical skills, and critical reflection. The 3wAI Framework by Bozkurt (2024b) addresses this need, focusing on "Know What," "Know How," and "Know Why" and aiming to promote foundational understanding, practical application, and ethical awareness. Zhao et al. (2024) asserts that GenAI literacy should include pragmatic, safety, reflective, socio-ethical, and contextual elements. Scholars such as Lyu et al. (2024) and Annapureddy et al. (2024) emphasise the distinction between general AI literacy and the specific skills required for GenAI, pointing out that existing frameworks often overlook the skills necessary for effectively utilising these tools.

The need for GenAI literacy is particularly crucial in educational settings. The absence of a comprehensive GenAI literacy framework poses challenges to its effective integration into learning environments (Annapureddy et al., 2024). Chiu (2024) underlines the necessity of empirically evaluating pedagogies that incorporate GenAI to determine their impact on student outcomes. GenAI literacy can significantly enhance language learning, as suggested by Alzubi (2024), while Bozkurt (2023) advocates for its inclusion in curricula to prepare students for an AI-augmented future. Despite the potential of GenAI tools to improve student learning, Lyu et al. (2024) discovered that student-generated prompts often lack quality, highlighting a deficiency in necessary skills. This underscores the urgent need to develop GenAI literacy so that students can fully harness the potential of these technologies.

2.2. AI literacy instruments

Multiple AI literacy measurement instruments have been developed to address various contexts, audiences, and facets of AI literacy. These facets encompass technical, ethical, behavioural, and contextual elements, reflecting the multifaceted nature of AI literacy. The Scale for the Assessment of Non-Experts' AI Literacy (SNAIL), created by Laupichler et al. (2023a), evaluates technical knowledge, critical analysis, and practical application of AI. This scale's validity was confirmed through factor analyses and a Delphi study (Laupichler et al., 2023b). The AI Literacy Scale (AILS) focuses on general AI literacy by measuring awareness, usage, evaluation, and ethics and was validated by subject matter experts (Wang et al., 2023). For specific audiences, the Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS) targets medical students, assessing cognition, ability, vision, and ethics (Karaca et al., 2021). Another survey, based on the Unified Theory of Acceptance and Use of Technology (UTAUT) and the Technological Pedagogical and Content Knowledge (TPACK) framework, examines pedagogical knowledge and AI use intentions among EFL teachers (An et al., 2023). Instruments for younger audiences include the AI Literacy Questionnaire (AILQ) by Ng et al. (2024), designed for secondary students, which assesses affective, behavioural, cognitive, and ethical dimensions. Chai et al.'s (2021) AI Literacy Instrument explores students' confidence, readiness, and perceptions of AI. For broader competencies, Carolus et al.'s (2023) Meta AI Literacy Scale (MAILS) covers ethics, persuasion literacy, and emotion regulation. Additionally, Pinski et al.'s (2023) AI Literacy Instrument targets AI professionals, focusing on human-AI interaction, AI processing, and task knowledge, with validation for reliability and robustness. Lastly, Lee and Park's (2024) ChatGPT Literacy Tool is currently the only instrument specifically designed to assess GenAI skills among university students, though it relies on self-reported data.

Despite progress in AI literacy assessment, substantial gaps persist. Most current instruments rely heavily on self-reported assessments, with minimal use of performance-based measurements. A systematic literature review by Lintner (2024) identified 13 self-reported and only three performance-based instruments, highlighting the predominant reliance on self-reported tools and the urgent need for more reliable assessments. While there are some performance-based measures for general AI literacy, such as Hornsberger's (2023) test, which includes 30 multiple-choice questions and a sorting item, and Chiu's (2024) test consisting of 25 multiple-choice questions, no such measures have been developed for GenAI literacy. Performance-based evaluations are crucial for GenAI literacy, given the importance of practical engagement and iterative interactions (Lintner, 2024; Laupichler et al., 2023b; Yan et al., 2024a). To address this gap, it is essential to develop new GenAI literacy instruments specifically targeting GenAI skills and incorporating performance-based assessments for a more accurate evaluation of learners' competencies. These instruments would help educators better understand how students interact with generative models, identify areas needing additional training, and design effective interventions to enhance students' abilities to use GenAI technologies. Apart from recent efforts to assess GenAI literacy using instruments such as Lee and Park's (2024) ChatGPT Literacy Scale (a self-reported measure), rigorous performance-based assessments for GenAI literacy remain largely absent in literature. Performance-based measures are necessary because self-reporting can introduce biases and inaccuracies, particularly when assessing relatively new competencies such as GenAI literacy. Consequently, a valid, reliable performance-based assessment explicitly designed for evaluating learners' actual abilities to understand and utilise GenAI tools effectively remains an unmet need in educational research and practice.

2.3. Classical test theory and item response theory

Classical Test Theory (CTT) and Item Response Theory (IRT) offer complementary methodological approaches essential for developing robust and reliable assessment instruments in educational research (De Champlain, 2010; Thorndike et al., 1991; Hambleton & Jones, 1993). These theories provide the foundation for evaluating the structural validity and reliability of assessment tools, ensuring they accurately measure intended constructs such as GenAI literacy (Thorndike et al., 1991; American Educational Research Association et al., 2014; De Champlain, 2010). Specifically, CTT is grounded in the principle that an individual's observed test score is a combination of a true score and an error score. It provides a straightforward framework for analysing test data, focusing primarily on the reliability of test scores and the consistency of test items. Reliability in CTT is often assessed using measures such as Cronbach's alpha, which evaluates how well the items within a test measure the same construct (Miller, 1995). CTT is instrumental in the initial stages of performance-based assessment development, ensuring that selected items reflect the latent constructs of interest, such as GenAI literacy (De Champlain, 2010). On the other hand, IRT offers a more sophisticated analysis by examining the relationship between an individual's ability and the probability of correctly responding to a test item. IRT provides detailed item-level information necessary for refining assessments, offering insights into how items function across varying levels of learner ability (American Educational Research Association et al., 2014; De Champlain, 2010). This approach allows for a comprehensive analysis of item characteristics such as difficulty and discrimination, ensuring that the test measures a wide range of skills effectively across diverse populations. Together, CTT and IRT provide a comprehensive framework that is crucial for the development and refinement of educational assessments.

2.4. External validity and domain knowledge

External validity is a crucial aspect of educational and psychological assessments, ensuring that the constructs measured by an instrument, such as GenAI literacy, can predict relevant learning outcomes and performances in varied contexts (Messick, 1995; Boateng et al., 2018). This concept is emphasised in Messick's unified theory of validity, which posits that the validation process must consider not only how well an instrument measures the intended construct but also how well the construct aligns with real-world tasks and external criteria (Messick, 1995). In the context of GenAI literacy, external validity involves the instrument's ability to accurately predict students' capacity to engage with and perform learning tasks using GenAI tools effectively. This predictive capability is critical, as the ultimate goal of assessing GenAI literacy is to ensure that learners are equipped with the skills necessary to apply these technologies in authentic educational and professional settings (Yan et al., 2024a; Annapureddy et al., 2024).

Domain knowledge plays a vital role in assessing the external validity of GenAI literacy tools (Alexander, 1992; Alexander & Judy, 1988). It is essential to consider and control for the varying levels of domain knowledge possessed by students (Alexander & Judy, 1988), as this can significantly influence how effectively they can utilise GenAI technologies within different learning tasks (Tricot & Sweller, 2014). Failure to account for domain knowledge could result in skewed assessments (Alexander, 1992), where the lack of subject-specific understanding might be mistaken for deficiencies in GenAI literacy itself. Consequently, assessing GenAI literacy necessitates a nuanced approach that considers the interplay between domain knowledge and the specific skills required to interact with GenAI tools.

2.5. Generative AI literacy assessment test (GLAT)

Building on these theoretical foundations, this study introduces the development and validation of the GLAT. Designed to measure GenAI literacy among higher education students, the GLAT stands on the robust foundations of psychological and educational measurement practices (Thorndike et al., 1991; American Educational Research Association et al., 2014). These foundations ensure that the GLAT can effectively assess key aspects of GenAI literacy, including foundational knowledge, application, ethical awareness, and critical evaluation capabilities (Annapureddy et al., 2024; Long & Magerko, 2020). Specifically, the following research questions were investigated, aiming to assess the validity and reliability of the GLAT in capturing learners' GenAI literacy and predicting their learning performance:

- RQ1: To what extent does the GLAT exhibit structural validity and reliability through classical test theory and item response theory?
- RQ2: To what extent does the GLAT measure of GenAI literacy demonstrate external validity by predicting learners' performance in learning tasks with GenAI chatbots compared to self-reported instruments?

3. Methods

The GLAT was developed following the established test development procedures outlined in Psychological and Educational Measurement (Thorndike et al., 1991). The development process involved: 1) creating a blueprint of relevant GenAI concepts, 2) generating an initial set of test items based on this blueprint, and 3) evaluating face and content validity through expert reviews and pilot studies, respectively. An item analysis was conducted using CTT, which involved selecting items based on item difficulty and discrimination index. The structural validity and reliability of the GLAT (RQ1) were assessed using IRT. The external validity of the GLAT (RQ2) was evaluated by analysing its effectiveness in predicting learners' performance on tasks involving interaction with

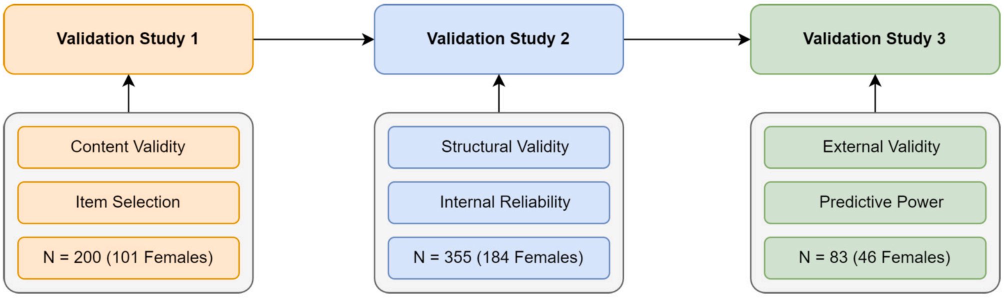

Fig. 1. The participant sample size and focus of each validation study.

Table 1
Dimensions and concepts related to generative AI.

Dimension	Concepts
Know & Understand	Generative AI (GenAI), Foundation model (e.g., LLM and diffusion model), Generation capability, Zero-shot learning, Prompt-based development, Content generation, Token, Artificial General Intelligence (AGI), Model alignment, RAG (Retrieval-Augmented Generation)
Use & Apply	Contextual understanding in content creation, Knowledge updating and integration, Information retrieval and synthesis, Token management and limitation, Multimodal content generation
Evaluate & Create	Trustworthiness of LLM outputs, LLM knowledge cutoff, Information cross-check (Hallucination), GenAI-generated content authenticity, Voice cloning with GenAI
Ethics	Model biases, Black box issue in GenAI, Copyright issues in GenAI-generated content, Content safety, Privacy concerns with GenAI

a GenAI chatbot, compared to a self-reported GenAI literacy instrument (Lee & Park, 2024). Details were elaborated in the following sections.

3.1. Participants

Three samples of higher education students were involved in various validation processes for the GLAT, following established standards (Thorndike et al., 1991; American Educational Research Association et al., 2014). As shown in Fig. 1, the first validation study involved assessing the content validity and selecting the GLAT item using CTT. Responses were gathered from 200 higher education students (101 females). The second validation study aimed to evaluate the structural validity and reliability of the GLAT using IRT (RQ1). This study analysed responses from 355 higher education students (184 females). The final validation study focused on assessing the external validity of the GLAT (RQ2) and included 83 higher education students (46 females). All participants were recruited through Prolific,[1] a reputable online research recruitment platform. Participants in the first and second studies received £1.5 for their time, while those in the final study were compensated £8 due to its increased complexity, as elaborated in Section 3.4. All studies were conducted using Qualtrics, with the item and option orders in the GLAT randomised (American Educational Research Association et al., 2014). Ethics approval was obtained from Monash University (Project ID: 37307), and informed consent was obtained from all participants.

3.2. Item generation

3.2.1. Blueprint construction

The blueprint of the GLAT was developed based on the four dimensions of AI literacy proposed by Ng et al. (2021b), including 1) Know & Understand, 2) Use & Apply, 3) Evaluate & Create, and 4) Ethics, and focusing on the specific context of GenAI. To identify a set of relevant GenAI concepts in each dimension, we cover a wide range of resources, including academic publications in prestige journals (e.g., Nature and Science), reports and articles published by reputable organisations (e.g., UNESCO, MIT News, and Standford HAI), and education information released by leading GenAI technology companies (e.g., OpenAI, Google, Meta, and NVIDIA). The decision to include diverse sources, beyond traditional academic publications, was driven by the rapidly evolving landscape of GenAI. This encompasses foundational models (e.g., large language models and diffusion models) as well as supporting infrastructure and techniques (e.g., embedding databases and retrieval methods). Our process for extracting relevant concepts involved three steps. Initially, two researchers independently reviewed the source documents ($n = 19$; links are available in the repository) and recorded pertinent GenAI concepts. Subsequently, they collaborated to consolidate similar concepts, resulting in a refined set of 25 concepts (see Table 1). A validation panel of three GenAI researchers then reviewed these concepts to ensure they provided reasonable coverage of the latest developments in GenAI.

3.2.2. Item generation

The item generation process began with aligning each item to the specific GenAI literacy concepts outlined in the blueprint (Table 1). This ensured comprehensive coverage across all dimensions of GenAI literacy, including knowledge, application, evaluation, and ethics. Each item was crafted as a multiple-choice question (MCQ) to assess understanding through a consistent and structured format, following the established guidelines (Haladyna, 2004). MCQs are particularly effective for evaluating knowledge across large participant groups due to their standardised nature and ease of scoring. A critical component of MCQ design is the creation of plausible distractors – incorrect answer options that are designed to challenge and differentiate between varying levels of participant understanding (Haladyna et al., 2002). Each distractor was carefully developed to reflect common misconceptions or logical errors relevant to the GenAI concept being assessed, ensuring they were plausible enough to create meaningful distinctions in responses. The drafting process involved multiple revisions to enhance the clarity, relevance, and cognitive demand of each question. Two researchers iteratively refined and checked the questions for ambiguity and ensured the options were free of overlapping meanings or unintended cues. An expert panel comprising specialists in GenAI, educational psychology, and psychometrics also reviewed each item and provided feedback for improvements. Specifically, our expert panel consisted of two experts: one GenAI researcher with extensive technical expertise and one educational psychologist and psychometrician specialising in educational assessment

[1] https://www.prolific.co/.

and measurement validation. Each expert panel member independently reviewed the items for clarity, appropriateness, and alignment with the instrument's four dimensions. Subsequent joint discussions were conducted to reach consensus on necessary modifications. This iterative process ensured clarity, content relevance, and cognitive appropriateness of the GLAT items. Eventually, the initial 25-item version of the GLAT was developed and assessed for content validity (Table 2).

3.2.3. Content validity

The content validity of the GLAT was assessed in a pilot study with a sample of 200 higher education students through six questions (Table 3) based on the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014). These questions assessed content validity from multiple aspects (Cronbach's alpha = 0.81), including relevance (C1), comprehensiveness (C2 and C3), comprehensibility (C4 and C5), and face validity (C6). Each question was measured using a five-point Likert scale, ranging from *strongly disagree* (1) to *strongly agree* (5). As shown in Table 3, the GLAT demonstrated strong content validity in all four aspects based on the pilot study results. This indicates that the GLAT is a valid tool for assessing GenAI literacy, as it effectively covers the necessary content areas, is easy to understand, and is perceived as an effective assessment tool by the students. This strong content validity sets the stage for the next phase of test development: ensuring structural validity and reliability through rigorous item selection.

3.3. RQ1: structural validity and reliability

3.3.1. Item selection

The item selection process utilised CTT, focusing on two key metrics: item difficulty and the discrimination index (Hambleton & Jones, 1993). Item difficulty was determined as the proportion of participants who answered an item correctly, with values ranging from 0 to 1. The discrimination index, quantified by the point-biserial correlation, reflects an item's ability to differentiate between high and low performers on the test. To calculate the discrimination index, we subtracted the number of test-takers in the lower group who answered the item correctly from the number of test-takers in the upper group who did so, then divided the result by the total number of test-takers. This index ranges from -1 to 1. As the GLAT aims to provide a continuous measure of GenAI literacy across various proficiency levels, we reported item difficulty without establishing a specific criterion (De Champlain, 2010). However, items with a discrimination index below 0.3 were excluded to ensure that the final set of items effectively differentiated among test-takers (Oosterhof, 2001). This systematic approach to evaluating and excluding test items based on discrimination indices is a standard method recommended broadly in psychometric literature to maintain the quality of educational assessment instruments (Crocker & Algina, 1986; Thorndike et al., 1991).

3.3.2. Structural validity

IRT models. After eliminating items with low discrimination indices, we assessed the structural validity of the final item set based on IRT (Reise & Waller, 2009). Three different IRT models were used: the Rasch model, the 2-parameter logistic (2PL) model, and the 3-parameter logistic (3PL) model. These models provide a nuanced understanding of the relationship between item characteristics and the latent trait being measured, allowing for an examination of each item's difficulty (b-parameter), discrimination (a-parameter), and guessing (c-parameter). Specifically, the Rasch model assumes that all items have the same discrimination and that guessing is not a factor, focusing solely on item difficulty. The 2PL model extends this by allowing each item to have its own discrimination parameter, which can help capture variations in how well different items differentiate between test-takers of different ability levels. The 3PL model further includes a guessing parameter (25% for four options MCQs), acknowledging that respondents may have a chance of answering an item correctly by guessing, especially in multiple-choice formats. The most appropriate IRT model was selected based on a comprehensive evaluation of model and item fit indexes, further elaborated below.

Assumption test. Before fitting IRT the models and evaluating model fit, a confirmatory factor analysis (single-factor model) was performed using the *lavaan* package in R (Rosseel, 2012) to assess the assumption of unidimensionality (Reise & Waller, 2009). This analysis verifies whether a single latent construct, GenAI literacy, could adequately explain all items in the GLAT. The model was considered unidimensional if it met the following criteria: $\chi^2/df < 2$, root mean square error of approximation (RMSEA) < 0.05, and standardized root mean square residual (SRMSR) < 0.1 (Brown, 2015). In addition, the assumption of local independence was evaluated using the Q3 statistic (Yen, 1984). A threshold value of 0.2 was applied (Chen & Thissen, 1997), with any pair of residual correlations exceeding this level indicating a potential violation of local independence.

Model fit. To evaluate model fit, the three IRT models were fitted using the *mirt* package in R (Chalmers, 2012), we utilised several fit statistics commonly used in comparing IRT models (Reise & Waller, 2009). The primary statistic used for model comparison was the likelihood ratio test, conducted using pairwise comparisons within an analysis of variance (ANOVA) framework, an approach widely recommended in psychometric literature (Reise & Waller, 2009) Specifically, we compared the simpler Rasch model to the more complex 2PL model, and then the 2PL model to the 3PL model. The ANOVA test provides insight into whether the increased complexity of a model significantly improves the fit of the data. This involves comparing the deviance (twice the negative log-likelihood) of each model, with a significant p-value indicating that the extra parameters provide a better fit. Alongside the likelihood ratio tests, we further assessed model fit using information criteria, including the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), to mitigate potential Type I errors from multiple statistical comparisons and ensure a robust model evaluation. These criteria take into account both the goodness of fit and the complexity of the model, with lower values suggesting a more preferable model. Additionally, we evaluated overall model fit with the M2 Statistic, RMSEA with values less than 0.06 indicating good fit, SRMSR with values below 0.08 suggesting good fit, and Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI) with values above 0.90 indicating acceptable fit and above 0.95 suggesting excellent fit (Reise & Waller, 2009; Maydeu-Olivares, 2013).

Item characteristic curves (ICCs). ICCs were plotted for each model to assess each item's ability to capture the latent trait of GenAI literacy consistently across varying proficiency levels (Reise & Waller, 2009).

Item fit. The signed chi-squared (S-χ^2) statistic was used to evaluate the item fit for each model (Orlando & Thissen, 2000). This method assesses how well the data fit the expected model by examining the extent of the difference between observed and expected response patterns for each item. To account for multiple tests and control the false discovery rate, the Benjamini-Hochberg procedure was applied to adjust the p-values (Benjamini & Hochberg, 1995).

3.3.3. Reliability evaluation

The reliability of the final item set was evaluated by focusing on internal consistency using Cronbach's alpha. Specifically, the internal consistency and reliability analyses were conducted on the final item set based on the selected, best-fitting 2PL IRT model, chosen after rigorous comparisons between Rasch, 2PL, and 3PL models. Cronbach's alpha provides an estimate of the proportion of variance in the test scores attributable to the true score variance. A Cronbach's alpha value of 0.7 or higher was deemed indicative of acceptable reliability in educational and psychological testing contexts, suggesting that the test items consistently assess the underlying construct of GenAI literacy (Miller, 1995). In addition, we utilised the coefficient omega, which is considered a

Table 2
The initial 25-item version of the Generative AI Literacy Assessment Test (GLAT).

Item	Dimension	Question	Options (Answer Highlighted)
1	Know & Understand	Which of the following best describes “Generative AI”?	**A. AI that creates new content like text, images, or music by learning from existing data.** B. An AI system designed to enhance the speed and accuracy of data retrieval in search engines. C. A form of artificial intelligence that focuses on translating languages in real-time. D. AI technology used primarily for managing and organizing large databases.
2	Know & Understand	Which of the following statements best describes an LLM (Large Language Model)?	A. It generates text by analyzing and summarizing large volumes of web content. **B. It generates text by predicting the next word based on the context of previous words.** C. It generates text by translating input text into multiple languages simultaneously. D. It generates text by using pre-defined templates and filling in the blanks.
3	Know & Understand	Which of the following tasks can Generative AI perform with a high degree of accuracy?	A. Predicting stock market trends B. Making ethical decisions in complex scenarios C. Diagnosing rare diseases **D. Generating human-like text based on prompts**
4	Know & Understand	In the context of Generative AI, what is “zero-shot learning”?	A. Training a model without any data. **B. The ability of a model to perform a task without any task-specific training.** C. A method of reducing the model’s training time to zero. D. A technique for generating synthetic training data.
5	Know & Understand	Which of the following is a potential challenge when using prompt-based development for text generation?	A. The language model can only generate binary outputs. B. The need for extensive labelled data to train the model. **C. Crafting a prompt that accurately captures the desired context and nuances.** D. The requirement for complex feature engineering.
6	Know & Understand	**[DROPPED]** When using a generative AI model to classify text into multiple categories, what is a common approach to handle more than two output classes?	A. Use multiple binary classifiers for each category. **B. Use a single prompt that includes all possible categories.** C. Train a separate model for each category. D. Use unsupervised learning to cluster the text data.
7	Know & Understand	What does the term “token” refer to in the context of a large language model (LLM)?	**A. A token is a unit of text, such as a word or a subword, that the model processes individually.** B. A token is a unique identifier assigned to each user interacting with the language model. C. A token is a security measure used to authenticate API requests to the language model. D. A token is a reward given to users for contributing valuable data to train the language model.
8	Know & Understand	Which of the following is NOT a requirement for an AI to be considered artificial general intelligence (AGI)?	A. The ability to learn and adapt to new tasks without human intervention. B. The capability to perform tasks across various domains with human-like proficiency. **C. The ability to predict future events with perfect accuracy.** D. The capacity to understand and generate natural language.
9	Know & Understand	**[DROPPED]** Why is model alignment important in the development of generative AI?	**A. To ensure AI systems better reflect human values and are safer.** B. To improve computational efficiency and reduce energy consumption. C. To enhance the alignment between model responses and user requests. D. To increase the speed of data processing and analysis.
10	Know & Understand	How does RAG (Retrieval-Augmented Generation) enhance the capabilities of an LLM?	A. By improving its grammar and syntax. **B. By providing it with real-time and relevant data.** C. By increasing its computational speed. D. By enabling it to understand multiple languages.
11	Use & Apply	When using generative AI to create a marketing pitch, which of the following strategies is least likely to be effective?	A. Supplying the AI with information about the target audience B. Asking the AI to include unique selling points and benefits C. Requesting the AI to use persuasive language techniques **D. Providing the AI with a list of competitors’ products**
12	Use & Apply	After deploying a customer service chatbot, you notice that it frequently provides outdated information about company policies. What is the best course of action to address this issue?	A. Implement a feedback loop where users can flag outdated information for review. **B. Schedule regular updates to the chatbot’s training data to include the latest company policies.** C. Set up a system where complex or policy-related queries are escalated to human agents for accurate responses. D. Conduct a comprehensive audit of the chatbot’s performance metrics to identify areas for improvement.
13	Use & Apply	Suppose you have a large dataset of emails and you want to build an application to answer questions based on this dataset. Which of the following scenarios best illustrates the advantage of using RAG over prompting (i.e., without RAG)?	A. You need to generate creative writing pieces based on the email content. B. You want to ensure the model can answer questions even if it has never seen similar questions before. **C. You need to answer questions that require specific information from different parts of the email dataset.** D. You want to reduce the size of the language model to save computational resources.
14	Use & Apply	**[DROPPED]** While using a Generative AI tool to write a story, you notice that the context window is limited to 500 tokens. What is a potential consequence of exceeding this limit?	A. The AI will automatically expand the context window B. The AI will ignore the excess tokens and generate text based on the first 500 tokens **C. The AI will generate text based on the most recent 500 tokens** D. The AI will stop functioning until the context window is reduced
15	Use & Apply	**[DROPPED]** When creating a video with a generative AI tool that supports text, images, and audio narration, which feature is most critical for ensuring the tool can handle this task effectively?	**A. Text-to-Speech (TTS) capability.** B. Image recognition capability. C. Multilingual support. D. Sentiment analysis capability.
16	Evaluate & Create	As a student using a Large Language Model (LLM) to gather information for an assignment, how should you approach the information it provides?	A. The LLM’s answers are always more trustworthy than any information you will find on the internet, so you can use them without further verification. B. The LLM’s answers are generally more trustworthy than internet sources, but you should still verify the information with other reliable sources. **C. The LLM’s answers are not necessarily more trustworthy than internet sources, and you should cross-check the information with other credible references.** D. The LLM’s answers are less trustworthy than internet sources because it relies on outdated information.

Table 2 (*continued*)

Item	Dimension	Question	Options (Answer Highlighted)
17	Evaluate & Create	It is unlikely for an LLM to provide an accurate summary of the latest financial market trends in real-time. Is this statement true or false?	**A. True, because the LLM's data may be outdated due to its knowledge cutoff.** B. True, because the LLM is not good at handling numbers and structured data. C. False, because the LLM frequently updates its knowledge base. D. False, because the LLM is capable of synthesizing the latest market data automatically.
18	Evaluate & Create	A generative AI tool has provided a summary of a research paper. The summary states, "The study found that increased screen time is directly correlated with decreased attention spans in children aged 8-12." What is your next step?	A. Accept the summary as accurate because AI tools are generally reliable. B. Ask the AI to provide more details about the study's methodology and results. **C. Cross-check the summary with the original research paper.** D. Use another AI tool to generate a summary for comparison and evaluate the consistency between both summaries
19	Evaluate & Create	While reviewing a video of a well-known public figure making controversial statements, which characteristic confirms the video was NOT generated by AI?	A. The public figure's voice sounds like themselves. B. The video has a professional and polished appearance. C. The video is high-quality with smooth transitions. **D. None of the above.**
20	Evaluate & Create	**[DROPPED]** Imagine someone is trying to clone your voice using generative AI. Which of the following audio recordings would be most effective?	A. A 1-hour recording of your conversation in a noisy public space. **B. A 30-second recording of your presentation speech.** C. A 5-minute recording of you singing a pop song. D. Generative AI cannot be used to clone human voice.
21	Ethics	When a generative AI system is used for screening job applications, what issue might arise concerning the quality and fairness of hiring decisions?	A. The AI system might overlook applicants' unique achievements and extracurricular activities. B. The AI system could misinterpret minor formatting differences in resumes. C. The AI system might not effectively handle applications submitted in various languages. **D. The AI system could reinforce existing biases found in historical hiring data.**
22	Ethics	In a healthcare startup, an accurate AI model recommends treatments, but doctors don't trust it because they can't understand how the model arrived at its conclusions. What core issue does this scenario illustrate?	A. The AI model uses obsolete training data. B. The training dataset lacks sufficient diversity. C. The treatment guidelines input are incorrect. **D. The AI model behaves as a black box.**
23	Ethics	What are the potential copyright implications for a journalist using an AI-generated image in a commercial article?	**A. The journalist needs to check the licensing policy of the AI tool they used.** B. The AI-generated image is automatically free to use without any restrictions. C. The journalist must pay a standard licensing fee to use the AI-generated image. D. The image cannot be used in any commercial context because it is AI-generated.
24	Ethics	Should we impose restrictions on the outputs of generative AI technologies?	A. Yes, to reduce the computational resources required for operating these technologies. **B. Yes, to prevent the dissemination of harmful or misleading content.** C. No, as it would hinder technological innovation and creativity. D. No, because users should have the freedom to access all generated content.
25	Ethics	Sending personal information to cloud-based generative AI tools has little privacy concerns.	A. True, as this information is encrypted using sophisticated algorithms during the transmission process. B. True, as generative AI tools are black-box systems and cannot output personal information even if it is used for model training. **C. False, as generative AI tools train on unencrypted data and can output private information based on their probabilistic nature.** D. False, as advancements in quantum computing can easily decipher the encrypted data.

Table 3
Content validity questions for the Generative AI Literacy Assessment Test (GLAT).

Question	Detail	Mean	SD
C1	The questions are directly related to generative AI concepts and skills.	4.62	0.65
C2	The test covers a broad range of concepts that are necessary for assessing generative AI literacy.	4.35	0.72
C3	The test includes questions that assess both fundamental and advanced concepts.	4.46	0.74
C4	The questions are clearly written and easy to understand.	4.26	0.79
C5	The options provided for each question are clearly distinct and easily distinguishable.	4.28	0.85
C6	Overall, I believe the test is an effective tool for assessing generative AI literacy.	4.22	0.87

more recent and potentially more accurate measure of internal consistency, particularly when items have varying loadings on the construct (Dunn et al., 2014). Specifically, we used omega total to evaluate the overall reliability of the GLAT in measuring GenAI literacy. This coefficient, like Cronbach's alpha, ranges from 0 to 1 and utilises a similar threshold value (e.g., 0.7) to indicate acceptable reliability. Additionally, to further understand and confirm the precision and reliability of the GLAT, we evaluated the test information function (Reise & Waller, 2009). The test information function was analysed to ensure consistent measurement at proficiency levels most relevant to the GLAT's intended application, particularly for learners and educators with low to average levels of GenAI literacy, considering GenAI technology is relatively new for them and in higher education (Annapureddy et al., 2024; Jin et al., 2024).

3.4. RQ2: external validity

The external validity of the GLAT was assessed by analysing its predictive power concerning learners' task performance during a task that involved interacting with a GenAI-powered conversational chatbot. The choice of a task involving interaction with a GenAI-powered chatbot is particularly appropriate for assessing external validity, as it closely mirrors real-world applications and challenges students might encounter in educational settings (McGrath et al., 2024), thereby providing a practical context for evaluating the external validity of GenAI literacy in pre-

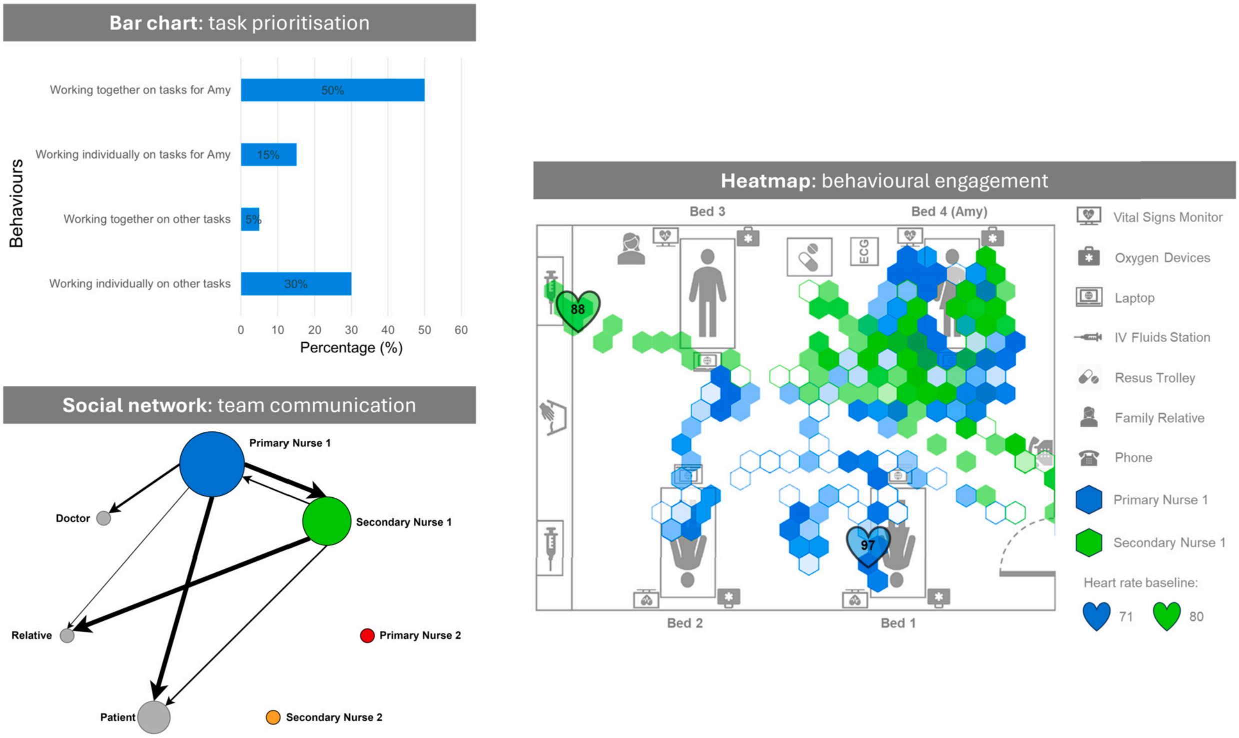

Fig. 2. Visual analytics on teamwork in healthcare simulations, including: a) a bar chart of four prioritisation strategies, b) a social network diagram of communication behaviours among the actors, and c) a ward map showing individuals' physical positions (hexagon), verbal communication duration (colour saturation), and peak heart rate locations.

dicting learner performance. This predictive capability was compared to that of a validated self-report instrument, the ChatGPT Literacy Scale (Lee & Park, 2024), to determine whether the GLAT provides additional predictive value beyond the self-report measure. The ChatGPT literacy scale serves as a suitable comparator, as it is specifically designed and validated for higher education students in the context of interactions with ChatGPT, a GenAI-powered conversational chatbot. Further details regarding the learning task, study procedure, and the analytical methods used are provided below.

3.4.1. *Learning task*

Learners engaged in a task that aimed to enhance their ability to comprehend complex visual analytics, an ability that many are lacking (Maltese et al., 2015; Donohoe & Costello, 2020), through interacting with GenAI chatbots. The learning task involved learners analysing a set of visual analytics on students' teamwork in healthcare simulations, composing a 100-word response and answering six evaluation questions to assess their ability to comprehend complex visual data. The visual analytics included three types of visualisations: a bar chart, a social network diagram, and a heatmap (Fig. 2). Specifically, the bar chart illustrated students' prioritisation strategies with positional data, simplifying the comparison of time spent on behaviours during the simulation (Yan et al., 2024c). The social network (sociogram) mapped interaction patterns via positional and audio data, highlighting communication frequencies and directions with the patient, doctor, and relative (Zhao et al., 2023). The advanced heatmap map combined students' physical positions, verbal durations, and peak heart rate locations. Inspired by sports analytics (Goldsberry, 2012), it used heatmaps to show verbal communication frequency and spatial distribution, and identified areas of peak physiological arousal.

A total of 83 higher education students (46 females) with medical, healthcare, and nursing backgrounds were involved in this validation to control for their familiarity with the healthcare simulation context. Learners were instructed to first analyse the visualisations and write a 100-word response on how the two nurses managed the primary patient, Amy, while attending to other beds, focusing on task prioritisation, verbal communication, and stress levels. After this, they answered six multiple-choice questions (two per visualisation) designed to assess comprehension of the visual data, addressing the first two levels of Bloom's taxonomy (knowledge and comprehension) (Bloom et al., 1984). For the knowledge questions, participants identified specific data points or patterns in visualisations, like determining which prioritisation behaviour two nurses spent the most time on from a bar chart. These questions assessed information retrieval skills. The comprehension questions required participants to interpret and derive insights, such as comparing spatial and verbal activities between two nurses using a ward map. These questions evaluated the ability to interpret insights and identify inconsistencies (see Table 4 for examples). Higher levels of Bloom's taxonomy, like application, were not considered because the learners have limited contextual knowledge of visual analytics.

Table 4
Example knowledge and comprehension question for the bar chart.

Bloom's Level	Question
Knowledge	Which behaviour did the two nurses spend the *most* time on?
Comprehension	How did the nurses spend their time working on tasks for Amy compared to other tasks?

3.4.2. *Chatbot design*

The GenAI chatbot was designed using the state-of-the-art retrieval-augmented generation (RAG) approach to improve response relevance and reduce inaccuracies (hallucinations) (Siriwardhana et al., 2023). As illustrated in Fig. 3, when learners ask a question about the visual analytics, the chatbot first retrieves relevant contextual information. It does this by computing vector embeddings of the prompts and calculating the cosine similarity between these prompt embeddings and stored knowledge embeddings (Li et al., 2024). The retrieved information, along with learners' questions and chat history, is then sent to a generative AI, specifically GPT-4o, to generate responses. The conversation is sub-

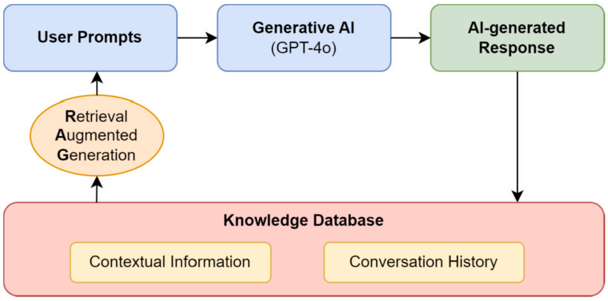

Fig. 3. System design of the generative AI (GenAI) chatbots.

sequently stored in the knowledge base to provide context for future interactions. This chatbot design is a representation of openly accessible GenAI chatbots, such as ChatGPT, Gemini, and Claude (Achiam et al., 2023).

3.4.3. Study procedure and measures

As illustrated in Fig. 4, learners begin by completing three literacy assessments: 1) the GLAT, 2) the ChatGPT Literacy Scale (Lee & Park, 2024), included as a comparative benchmark measure to evaluate the predictive power of the self-reported instrument, and 3) the mini-VLAT (Pandey & Ottley, 2023), which evaluates domain knowledge, specifically, learners' visualisation literacy pertinent to the task. Following a within-subject design, learners first perform the learning task independently, serving as the baseline condition. Subsequently, they repeat the task using different visual analytics (identical visualisations but with different data) while receiving support from the GenAI chatbot, constituting the AI-assisted condition. Consequently, five measures were captured and analysed: performance-based GenAI literacy using the GLAT (*GLAT-literacy*), self-reported GenAI literacy via the ChatGPT Literacy Scale (*ChatGPT-literacy*), visualisation literacy through the mini-VLAT (*VLAT-literacy*), baseline task performance without support (*baseline score*), and task performance with GenAI chatbot assistance (*AI-assisted score*). Task performance data was recorded via a website developed specifically for this study, capturing evaluation scores and written responses in both baseline and AI-assisted conditions, along with chatbot interactions in the AI-assisted condition.

3.4.4. Predictive analysis

Predictive modelling was conducted to evaluate the external validity of the GLAT in predicting task performance compared to the ChatGPT Literacy Scale. Specifically, the analysis was explicitly designed to examine how well GLAT scores could predict learners' ability to interact effectively with GenAI tools in realistic educational scenarios. Although the task involved an AI-assisted learning context, the primary goal was to validate the predictive strength of the GLAT itself, rather than to assess the educational effectiveness or impact of the AI-assisted intervention.

To achieve this, we used ordinary least squares (OLS) regression to examine the predictive power of learners' *GLAT-literacy* and *ChatGPT-literacy* (independent variables; IVs) on their *AI-assisted score* (dependent variable; DV), while controlling for their *baseline score* and *VLAT-literacy* (independent variables; IVs). Each measure was first standardised to ensure a uniform format for interpretation. Interaction terms were excluded as an ANOVA revealed no significant model improvements ($F(11,67) = 1.45, p = .17$). The final regression included an intercept (β_0) and a main effect for each IV (β_1 to β_4). All assumptions for the regression analysis were verified. Linearity was confirmed by plotting predicted versus observed values. The normality of residuals was checked using the Shapiro-Wilk test and QQ plots. Homoscedasticity was assessed with the Breusch-Pagan test, and the independence of residuals was evaluated using the Durbin-Watson test. All assumptions were satisfied.

Table 5
Item difficulty and discrimination indices for each item.

Item	Difficulty	Discriminative Index	Item	Difficulty	Discriminative Index
1	0.892	0.485	14	0.299	**0.272**
2	0.270	0.381	15	0.363	**0.205**
3	0.902	0.358	16	0.696	0.489
4	0.578	0.338	17	0.613	0.479
5	0.613	0.432	18	0.721	0.390
6	0.245	**0.064**	19	0.804	0.332
7	0.574	0.339	20	0.721	**0.230**
8	0.853	0.448	21	0.691	0.555
9	0.304	**0.249**	22	0.627	0.500
10	0.676	0.354	23	0.706	0.376
11	0.593	0.374	24	0.799	0.378
12	0.730	0.406	25	0.603	0.432
13	0.485	0.355			

$$\begin{aligned} \textit{AI-assisted score} = {} & \beta_0 + \beta_1 \times \textit{GLAT-literacy} + \beta_2 \times \textit{ChatGPT-literacy} \\ & + \beta_3 \times \textit{baseline score} + \beta_4 \times \textit{VLAT-literacy} \end{aligned} \tag{1}$$

4. Results

4.1. Structural validity and reliability (RQ1)

4.1.1. Item selection

The item selection process for the GenAI Literacy Assessment Test (GLAT) revealed varying degrees of item difficulty and discrimination indices that facilitated the identification of items for inclusion in the final assessment. As shown in Table 5, a total of five items (Items 6, 9, 14, 15, and 20) had discrimination indices below the threshold of 0.3. These items were consequently excluded from the final item set to ensure that the assessment effectively differentiates among test-takers. Specifically, Item 6 had a discrimination index of 0.06, Item 9 had a discrimination index of 0.25, Item 14 had a discrimination index of 0.27, Item 15 had a discrimination index of 0.21, and Item 20 had a discrimination index of 0.23. The remaining items demonstrated adequate discriminability, with indices ranging from 0.33 to 0.55 (M = 0.41, SD = 0.06), indicating that the retained items have a consistent ability to differentiate between high and low performers, with relatively low variability in their discriminative power. Of the 20 items retained, the item difficulties ranged between 0.25 and 0.90 (M = 0.67, SD = 0.14), indicating that, on average, the items tend to be moderately easy with a moderate spread in item difficulty. This spread ensures a diverse range of difficulty levels across the items, catering to various proficiency levels of the test-takers. Fig. 5 presents a descriptive summary of the GLAT scores obtained from the validation sample. The score distribution (M = 12.80; SD = 4.14) indicates a moderately symmetrical pattern, suggesting appropriate variability across participants and good coverage of different GenAI literacy levels in the validation sample.

4.1.2. Assumption evaluation

Both the assumptions of unidimensionality and local independence were confirmed. The single-factor model demonstrated a χ^2/df ratio of 1.51, which is below the recommended threshold of 2, indicating a good fit for the data. The RMSEA was 0.038, with a 90% confidence interval ranging from 0.028 to 0.048. This value is well below the criterion of 0.05, suggesting an excellent fit. Additionally, the SRMSR was 0.050, meeting the criterion of less than 0.10. Additionally, the examination of residual correlations revealed that no pairs exceeded the threshold (0.2), indicating that the assumption of local independence was confirmed.

4.1.3. Structural validity

We sequentially fitted three progressively complex IRT models (the Rasch, 2PL, and 3PL models) to evaluate whether increased complexity led to meaningful improvements in modelling GLAT data, with the aim of selecting the most appropriate model that balanced parsimony

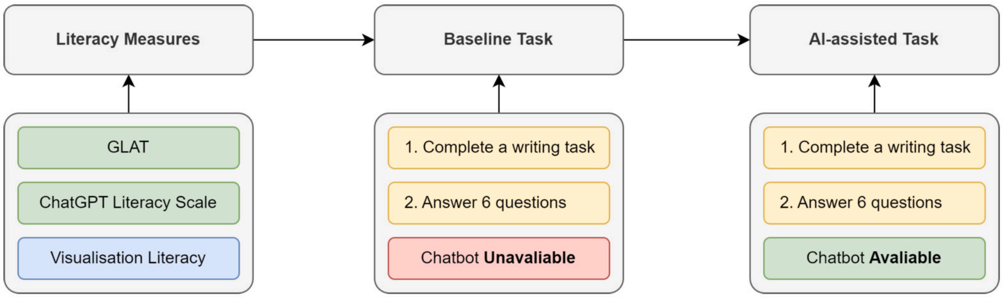

Fig. 4. Study design: three literacy measurements, a baseline task, and an AI-assisted task.

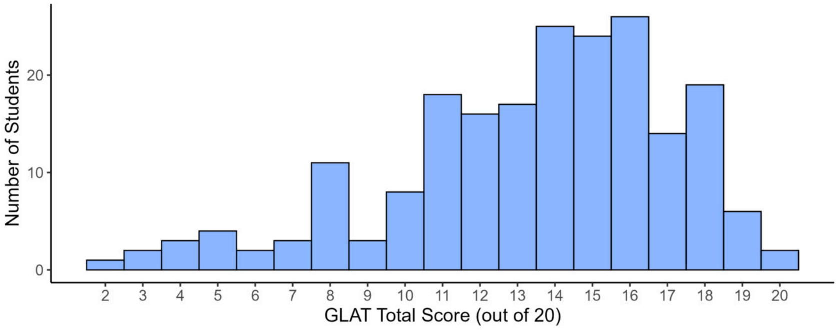

Fig. 5. Distribution of GLAT score on the validation sample.

Table 6
Fit indices and information criteria for the Rasch, 2PL, and 3PL models.

Model	M2	RMSEA	SRMR	TLI	CFI	AIC	BIC
Rasch	292.340	0.040	0.072	0.944	0.945	7823.532	7904.246
2PL	225.325	0.031	0.052	0.967	0.970	7805.936	7959.678
3PL	166.014	0.018	0.051	0.989	0.991	7822.048	8052.660

with goodness of fit. For the comparison between the Rasch and the 2PL model, the ANOVA results indicated a significant improvement in fit with the 2PL model $\chi^2 = 55.596$, $df = 19$, $p < 0.001$, suggesting that the added complexity of allowing for varying item discriminations provided a better fit to the data. However, for the comparison between the 2PL and the 3PL model, the ANOVA results showed that the 3PL model did not significantly improve the fit over the 2PL model $\chi^2 = 23.888$, $df = 20$, $p = 0.247$, indicating that accounting for guessing parameters did not significantly enhance the model's explanatory power.

As shown in Fig. 6, the items in all three models demonstrated an S-curve shape, indicating that the probability of a correct response increases with higher levels of the latent trait. These patterns are consistent with expectations for assessments that measure proficiency like GenAI literacy (Reise & Waller, 2009). Furthermore, as shown in Table 7, the signed chi-squared item fit indices for the Rasch, 2PL, and 3PL models vary across the items. Items 4, 11, and 19 stand out with significant p-values, indicating potential model misfit under the Rasch model. Item 19 also shows a significant misfit under the 2PL model and the 3PL model. In contrast, other items generally exhibit non-significant p-values, suggesting an adequate fit for those items across different models. This item fit evaluation indicates that both 2PL and 3PL models may provide a better fit for the majority of items compared to the Rasch model.

Based on these analyses, the 2PL model was determined to be the best-fitting model. The fit indices and information criteria for the 2PL model indicated a robust structural validity (Table 6): the M2 statistic was 225.325, and the RMSEA was 0.031, indicating a good fit. The SRMSR was 0.052, which also suggests a good fit, while both the TLI and CFI were 0.967 and 0.970, respectively, suggesting an excellent fit. Additionally, the AIC and BIC were 7805.936 and 7959.678, respectively, supporting the 2PL model as more favourable when balancing fit and model complexity.

Table 7
Signed chi-squared item fit indices for the Rasch, 2PL, and 3PL models.

Item	Rasch			2-PL			3-PL		
	S-χ^2	df	p	S-χ^2	df	p	S-χ^2	df	p
1	24.164	13	.120	18.887	11	.252	18.134	10	.212
2	15.151	10	.212	14.322	10	.393	12.708	10	.536
3	19.574	13	.198	16.637	12	.393	15.409	11	.471
4	29.801	13	**.033**	19.875	15	.393	16.147	13	.536
5	21.876	13	.163	21.642	13	.252	22.385	12	.212
7	24.593	13	.120	14.831	15	.639	13.813	14	.714
8	8.871	13	.824	5.632	12	.974	6.141	11	.909
10	7.183	13	.892	5.671	14	.974	6.600	13	.922
11	30.991	13	**.030**	25.162	13	.220	27.214	13	.120
12	15.094	14	.499	14.616	14	.623	14.661	13	.585
13	18.716	12	.198	16.549	14	.518	15.418	11	.471
16	10.928	13	.686	10.023	13	.807	9.903	12	.805
17	13.555	11	.398	13.135	11	.518	7.847	11	.808
18	13.998	13	.499	14.315	13	.587	14.259	12	.568
19	43.512	14	<**.001**	39.537	14	<**.001**	32.063	13	**.040**
21	23.203	13	.130	21.786	12	.252	12.172	11	.585
22	20.463	13	.198	10.580	11	.639	8.435	10	.805
23	13.160	14	.642	10.484	14	.807	10.126	13	.805
24	12.111	14	.686	12.717	14	.686	10.509	13	.805
25	16.962	11	.198	17.441	11	.317	19.571	11	.212

4.1.4. Reliability

The reliability of the GLAT was assessed by examining internal consistency using Cronbach's alpha and omega total. Cronbach's alpha was calculated to be 0.80, indicating good reliability, as it surpassed the threshold of 0.7 commonly used in educational testing (Miller, 1995). In addition, the coefficient omega total was computed to further evaluate reliability, resulting in a value of 0.81. This suggested a high level of internal consistency and confirmed the reliability of the GLAT in measuring GenAI literacy, particularly given the varying loadings on the construct (Dunn et al., 2014). Additionally, the precision and reliability of the GLAT were further examined using the test information function. As Fig. 7 shows, the GLAT provided the most information at a profi-

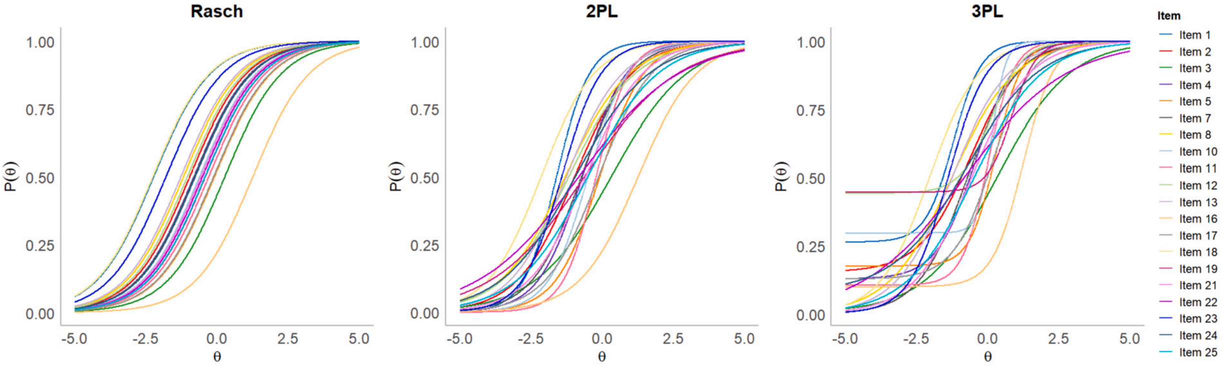

Fig. 6. Item characteristic curves for the Rasch, 2PL, and 3PL models. θ presents the latent trait, GenAI literacy.

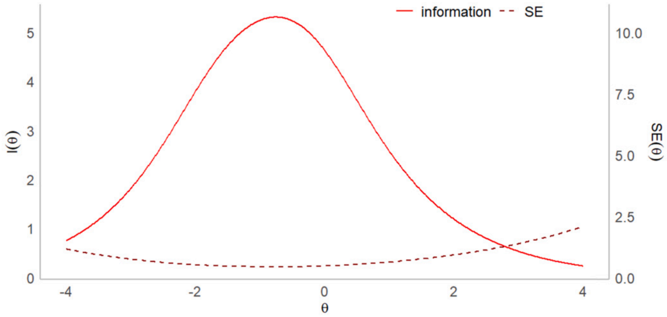

Fig. 7. Test information function for the 2PL models. θ presents the latent trait, GenAI literacy.

ciency level of $\theta = -0.8$, indicating that the test is highly reliable for individuals with below-average GenAI literacy. The maximum information value of 5.31 at this point suggests strong measurement precision for this target group. However, the information decreases for proficiency levels further from $\theta = -0.8$, especially for individuals with higher proficiency, where the test is less discriminative. The standard error (SE) was also lowest around $\theta = -0.8$, further supporting the test's high precision for low to moderate GenAI literacy, while precision diminished for more extreme proficiency levels.

4.2. External validity (RQ2)

The predictive model was statistically significant, $F(4, 78) = 9.233$, $p < .001$, with an R^2 of 0.321, indicating that the model accounted for approximately 32.1% of the variance in *AI-assisted score*. Among the predictors, *GLAT-literacy* ($\beta = 0.220$, $t = 2.093$, $p = .040$) and *VLAT-literacy* ($\beta = 0.322$, $t = 2.946$, $p = .004$) were both significant positive predictors of *AI-assisted score*. For a standard deviation increase in *GLAT-literacy*, the *AI-assisted score* increased by approximately 0.220 standard deviations, suggesting that greater proficiency in GenAI literacy is associated with enhanced performance in tasks supported by the GenAI chatbot. Similarly, a standard deviation increase in *VLAT-literacy* resulted in an increase of approximately 0.322 standard deviations in the *AI-assisted score*, indicating that domain knowledge, such as visualisation literacy, significantly contributes to improved task performance, which is expected considering the task involved comprehending visual analytics. Whereas, *ChatGPT-literacy* showed a negative relationship with the *AI-assisted score* ($\beta = -0.159$), but was not significant $t = -1.579$, $p = 0.118$). This suggests students' self-reported proficiency with ChatGPT was not a significant factor in predicting their performance scores in GenAI-assisted tasks. The baseline score ($\beta = 0.098$, $t = 0.907$, $p = 0.367$) did not significantly predict the *AI-assisted score*, suggesting that initial task performance without AI assistance did not substantially influence outcomes when using the GenAI chatbot. This highlights the independent contributions of *GLAT-literacy* and *VLAT-literacy* to learners' task performance.

5. Discussion

Effective and valid measurement of GenAI literacy is essential in higher education as learners and educators increasingly encounter GenAI tools in their study, work, and daily lives (Yan et al., 2024a; Cukurova, 2024; Khosravi et al., 2023). This study developed and validated the GLAT in line with established standards for psychological and educational measurement (Thorndike et al., 1991; American Educational Research Association et al., 2014). Regarding RQ1, the GLAT demonstrated a 2PL model with strong structural validity, meeting the requirements of item discrimination and difficulty across a diverse sample. This indicates that the GLAT effectively differentiates individuals with varying GenAI literacy levels, which is crucial for accurately assessing competencies related to GenAI use (Annapureddy et al., 2024; Yan et al., 2024a; Zhao et al., 2024; Bozkurt, 2024a). The GLAT also showed good reliability, particularly in assessing students with low to moderate GenAI literacy. This aligns with its intended use, considering the current state of GenAI literacy, where the technology is relatively new and integrated training is limited in higher education curricula (Holmes et al., 2023; Jin et al., 2024). The GLAT is thus especially valuable for identifying individuals who may need additional education and support to effectively understand and use GenAI ethically. These findings highlight the GLAT's utility in assessing foundational GenAI knowledge, particularly where students have limited prior exposure or training. However, as GenAI training becomes more integrated into higher education and students' average GenAI literacy improves, the GLAT will require updates to remain relevant. This aligns with the need for an iterative design process for test instruments to adapt to new data and evolving use contexts (American Educational Research Association et al., 2014).

In terms of RQ2 and external validity, we examined the extent to which the GLAT predicts learners' performance in tasks involving GenAI chatbots compared to self-reported instruments, using visualisation literacy and baseline performance as control variables. The predictive model showed that GenAI literacy, as measured by GLAT, was a significant predictor of learners' performance in GenAI-supported tasks, whereas domain knowledge (e.g., visualisation literacy) served as a control to account for differences in learners' comprehension of visual information. The significant positive relationship between GLAT and AI-assisted task performance underscores the value of reliably assessing GenAI literacy to predict real-world learner outcomes (Annapureddy et al., 2024; Chiu, 2024). In contrast, self-reported ChatGPT proficiency was not a significant predictor, highlighting the limitations of self-assessment, which may be prone to biases or inaccuracies (Lintner, 2024; Ng et al., 2021b). The control for domain knowledge (e.g., visualisation literacy) ensured that the observed effects were specific to GenAI literacy, thereby reinforcing the importance of targeted skill development in GenAI (Lee & Park, 2024; Lyu et al., 2024). These findings suggest that enhancing GenAI literacy has a direct effect on learners' ability to effectively engage with GenAI tools, independent of their domain knowledge. This insight is critical for educators aiming to design targeted interventions that bolster students' competencies in using GenAI technologies effectively in diverse educational contexts.

5.1. Implications to research and practice

The study's findings have profound implications for advancing research and practice in GenAI literacy within higher education. The development of the GLAT underscores the need for performance-based measures over traditional self-reported assessments, addressing the limitations and biases inherent in self-assessment tools (Ng et al., 2021b; Lintner, 2024). This shift to more reliable measures will enable educators and researchers to make informed decisions about integrating GenAI into educational settings. For educators, leveraging the GLAT offers a diagnostic tool to assess and enhance students' GenAI literacy, highlighting individual needs for targeted interventions (Alzubi, 2024; Bozkurt, 2023). The assessment's external validity in GenAI-supported learning tasks further underscores its practical utility, providing insights into students' preparedness to navigate GenAI technologies in diverse educational contexts. By incorporating the GLAT into curriculum development, educators can better align teaching strategies with the specific competencies required for effective GenAI engagement, thereby preparing students for an AI-driven future. Furthermore, the study encourages researchers to adopt a comprehensive, iterative approach to developing and validating educational assessments, ensuring their continued relevance amidst the evolving GenAI landscape (Holmes et al., 2023; Jin et al., 2024). By focusing on multidimensional literacy frameworks that integrate foundational knowledge, practical skills, and ethical understanding, future research can enhance the robustness of educational instruments and cultivate a nuanced understanding of GenAI literacy in academia (Zhao et al., 2024; Bozkurt, 2024a).

5.2. Limitations and future directions

While this study advances the measurement of GenAI literacy in higher education, several limitations warrant attention. Firstly, the GLAT was developed and validated primarily with higher education students, excluding younger K–12 students and educators. Additionally, the current GLAT items predominantly use specialised GenAI-related terminology, which may further restrict their direct applicability to other academic disciplines or more general educational settings. Future research should therefore extend the instrument's applicability across diverse educational levels, participant groups, and subject areas, adapting or expanding the GLAT as needed to ensure its broader relevance and utility. Furthermore, the study's examination of external validity is based on context-specific tasks involving visual analytics and chatbot interactions. Future investigations should incorporate various contexts and task complexities to comprehensively understand how GenAI literacy affects learning performance across different domains.

Another notable limitation is the focus on certain types of domain knowledge, such as visualisation literacy, without considering other relevant knowledge areas that may affect task performance. Thus, future studies should examine a wider range of domain knowledge to better control its influence on GenAI literacy outcomes. The rapidly evolving nature of GenAI technology presents another limitation. As these tools advance, so too must the instruments assessing GenAI literacy. Researchers should continuously update and refine the GLAT to keep pace with new developments and maintain its effectiveness and accuracy (American Educational Research Association et al., 2014). Lastly, it is important to note that the test is conducted in English, which may limit its accessibility and relevance for non-English-speaking participants. Future studies should explore the adaptation of the assessment for different languages to ensure its validity and applicability across diverse linguistic populations. In addition, integrating data-mining and visual analytics techniques can enable researchers to longitudinally track GLAT usage, allowing for a deeper exploration of how assessment score trajectories relate to authentic learning behaviours in real-world educational contexts.

6. Conclusion

This study introduced the GLAT, a performance-based instrument designed to assess GenAI literacy within higher education contexts. The GLAT demonstrated robust structural validity and reliability, particularly in evaluating foundational GenAI knowledge among students with varying levels of expertise. The external validity of the GLAT further underscored its practical utility, showcasing a significant positive relationship between GLAT scores and learners' performance in GenAI-supported tasks. This study advocates for the integration of performance-based assessments in addition to traditional self-reported measures to evaluate GenAI literacy reliably. The findings highlight the need for continuous adaptation of assessment tools to keep pace with technological advancements, thereby equipping educators and students with the skills necessary to engage in an AI-driven future effectively. Future research should focus on expanding the applicability of the GLAT across diverse educational levels and contexts, addressing the complex and evolving landscape of GenAI technologies.

CRediT authorship contribution statement

Yueqiao Jin: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Roberto Martinez-Maldonado:** Writing – review & editing, Conceptualization. **Dragan Gašević:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Lixiang Yan:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Statements on open data and ethics

The study was approved by an ethical committee with ID: 37307. Informed consent was obtained from all participants, and their privacy rights were strictly observed. The data can be obtained by sending request e-mails to the corresponding author.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.caeai.2025.100436.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint, arXiv:2303.08774.

Alexander, P. A. (1992). Domain knowledge: Evolving themes and emerging concerns. *Educational Psychologist*, *27*, 33–51.

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*, 375–404.

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of ai literacy conceptualization, constructs, and implementation and assessment efforts (2019-2023). *Computers and Education Open*, *100173*.

Alzubi, A. A. F. (2024). Generative artificial intelligence in the efl writing context: Students' literacy in perspective. *Qubahan Academic Journal*, *4*, 59–69.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

An, X., Chai, C. S., Li, Y., Zhou, Y., Shen, X., Zheng, C., & Chen, M. (2023). Modeling English teachers' behavioral intention to use artificial intelligence in middle schools. *Education and Information Technologies*, *28*, 5187–5208.

Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2024). Generative ai literacy: Twelve defining competencies. *Digital Government: Research and Practice*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological*, *57*, 289–300.

Bloom, B. S., Krathwohl, D. R., Masia, B. B., et al. (1984). *Bloom taxonomy of educational objectives*. London: Pearson Education.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149.

Bozkurt, A. (2023). *Unleashing the potential of generative ai, conversational agents and chatbots in educational praxis: A systematic review and bibliometric analysis of genai in education*.

Bozkurt, A. (2024a). Why generative ai literacy, why now and why it matters in the educational landscape? Kings, queens and genai dragons. *Open Praxis*, *16*, 283–290.

Bozkurt, A. (2024b). *Why generative ai literacy, why now and why it matters in the educational landscape? Kings, queens and genai dragons*.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Burgsteiner, H., Kandlhofer, M., & Steinbauer, G. (2016). Irobot: Teaching the basics of artificial intelligence in high schools. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1324–1337).

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). Mails-meta ai literacy scale: Development and testing of an ai literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans*, *1*, Article 100014.

Chai, C. S., Lin, P. Y., Jong, M. S. Y., Dai, Y., Chiu, T. K., & Qin, J. (2021). Perceptions of and behavioral intentions towards learning artificial intelligence in primary school students. *Educational Technology & Society*, *24*, 89–101.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Chiu, T. K. (2024). Future research recommendations for transforming higher education with generative ai. *Computers and Education: Artificial Intelligence*, *6*, Article 100197.

Chiu, T. K., Chen, Y., Yau, K. W., Chai, C. S., Meng, H., King, I., Wong, S., & Yam, Y. (2024). Developing and validating measures for ai literacy tests: From self-reported to objective measures. *Computers and Education: Artificial Intelligence*, *7*, Article 100282.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.

Cukurova, M. (2024). *The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence*. BJET.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*, 109–117.

Donohoe, D., & Costello, E. (2020). Data visualisation literacy in higher education: An exploratory study of understanding of a learning dashboard tool. *International Journal: Emerging Technologies in Learning*, *15*. https://doi.org/10.3991/ijet.v15i17.15041.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399–412.

Goldsberry, K. (2012). Courtvision: New visual and spatial analytics for the nba. In *2012 MIT sloan sports analytics conference* (pp. 12–15).

Haladyna, T. M. (2004). Developing and validating multiple-choice test items. *Routledge*.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–333.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, Issues and Practice*, *12*, 38–47.

Holmes, W., Miao, F., et al. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an ai literacy test. *Computers and Education: Artificial Intelligence*, *5*, Article 100165.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*, 1–38.

Jin, Y., Yan, L., Echeverria, V., Gašević, D., & Martinez-Maldonado, R. (2024). Generative ai in higher education: A global perspective of institutional adoption policies and guidelines. arXiv preprint, arXiv:2405.11800.

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE frontiers in education conference (FIE)* (pp. 1–9). IEEE.

Karaca, O., Çalışkan, S. A., & Demir, K. (2021). Medical artificial intelligence readiness scale for medical students (mairs-ms)–development, validity and reliability study. *BMC Medical Education*, *21*, 1–9.

Khosravi, H., Viberg, O., Kovanovic, V., & Ferguson, R. (2023). Generative ai and learning analytics. *Journal of Logic and Algebraic*, *10*, 1–6.

Koch, M. J., Wienrich, C., Straka, S., Latoschik, M. E., & Carolus, A. (2024). Overview and confirmatory and exploratory factor analysis of ai literacy scale. *Computers and Education: Artificial Intelligence*, *100310*.

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023a). Development of the "scale for the assessment of non-experts' ai literacy"–an exploratory factor analysis. *Computers in Human Behavior Reports*, *12*, Article 100338.

Laupichler, M. C., Aster, A., & Raupach, T. (2023b). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' ai literacy. *Computers and Education: Artificial Intelligence*, *4*, Article 100126.

Lee, S., & Park, G. (2024). Development and validation of chatgpt literacy scale. *Current Psychology*, 1–13.

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., & Dou, Z. (2024). From matching to generation: A survey on generative information retrieval. arXiv preprint, arXiv:2404.14851.

Lintner, T. (2024). A systematic review of ai literacy scales. *npj Science of Learning*, *9*. https://doi.org/10.1038/s41539-024-00264-4.

Long, D., & Magerko, B. (2020). What is ai literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–16). ACM.

Lyu, W., Wang, Y., Chung, T., Sun, Y., & Zhang, Y. (2024). Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In *Proceedings of the eleventh ACM conference on learning @ scale* (pp. 63–74).

Maltese, A. V., Harsh, J. A., & Svetina, D. (2015). Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *Journal of College Science Teaching*, *45*, 84–90. https://doi.org/10.2505/4/jcst15_045_01_84.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement Interdisciplinary Research & Perspective*, *11*, 71–101.

McDonald, N., Johri, A., Ali, A., & Hingle, A. (2024). Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines. arXiv preprint, arXiv:2402.01659.

McGrath, C., Farazouli, A., & Cerratto-Pargman, T. (2024). Generative ai chatbots in higher education: A review of an emerging research area. *Higher Education*, 1–17.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist*, *50*, 741–749.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling. A Multidisciplinary Journal*, *2*, 255–273. https://doi.org/10.1080/10705519509540013.

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021a). Ai literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*, 504–509.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021b). Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, Article 100041.

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the ai literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology*, *55*, 1082–1104.

Oosterhof, A. (2001). *Classroom applications of educational measurement*. ERIC.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

Pandey, S., & Ottley, A. (2023). Mini-vlat: A short and effective measure of visualization literacy. In *Computer graphics forum* (pp. 1–11). Wiley Online Library.

Pinski, M., & Benlian, A. (2023). Ai literacy-towards measuring human competency in artificial intelligence. In *Hawaii international conference on system sciences 2023 (HICSS-56)*. *3* (pp. 165–174).

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, *11*, 1–17.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. Macmillan Publishing Co, Inc.

Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, *26*, 265–283.

Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*, 1324–1337.

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024a). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*. Accepted for publication.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024b). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*, 90–112.

Yan, L., Echeverria, V., Jin, Y., Fernandez-Nieto, G., Zhao, L., Li, X., Alfredo, R., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024c). Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning. *British Journal of Educational Technology*, *55*(5), 1900–1925.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Zhao, L., Swiecki, Z., Gasevic, D., Yan, L., Dix, S., Jaggard, H., Wotherspoon, R., Osborne, A., Li, X., & Alfredo, R. (2023). METS: Multimodal learning analytics of embodied teamwork learning. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 186–196).

Zhao, X., Cox, A., & Cai, L. (2024). Chatgpt and the digitisation of writing. *Humanities & Social Sciences Communications*, *11*, 1–9.