Psicothema (2026) 38(1) 1-12

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Article

Using Artificial Intelligence in Test Construction: A Practical Guide

Javier Suárez-Álvarez[1], Qiwei He[2], Nigel Guenole[3] and Damiano D'Urso[4]

[1] University of Massachusetts Amherst (USA)
[2] Georgetown University (USA)
[3] University of London (United Kingdom)
[4] Independent Researcher (Netherlands)

ARTICLE INFO

Received: 01/08/2025
Accepted: 09/10/2025

Keywords:
Artificial intelligence
Test construction
Automated item generation
Validity
Fairness

ABSTRACT

Background: Artificial Intelligence (AI) is increasingly used to enhance traditional assessment practices by improving efficiency, reducing costs, and enabling greater scalability. However, its use has largely been confined to large corporations, with limited uptake by researchers and practitioners. This study aims to critically review current AI-based applications in test construction and propose practical guidelines to help maximize their benefits while addressing potential risks. **Method:** A comprehensive literature review was conducted to examine recent advances in AI-based test construction, focusing on item development and calibration, with real-world examples to demonstrate practical implementation. **Results:** Best practices for AI in test development are evolving, but responsible use requires ongoing human oversight. Effective AI-based item generation depends on quality training data, alignment with intended use, model comparison, and output validation. For calibration, essential steps include defining construct validity, applying prompt engineering, checking semantic alignment, conducting pseudo factor analysis, and evaluating model fit with exploratory methods. **Conclusions:** We propose a practical guide for using generative AI in test development and calibration, targeting challenges related to validity, reliability, and fairness by linking each issue to specific guidelines that promote responsible, effective implementation.

Uso de la Inteligencia Artificial en la Construcción de Pruebas: Una Guía Práctica

Palabras clave:
Inteligencia artificial
Construcción de pruebas
Generación automatizada de ítems
Validez
Equidad

RESUMEN

Antecedentes: La inteligencia artificial (IA) se utiliza crecientemente para mejorar las prácticas tradicionales de evaluación, aumentando la eficiencia, reduciendo costos y facilitando la escalabilidad. Sin embargo, su uso se ha limitado a grandes corporaciones, con escasa adopción por parte de investigadores y profesionales. Este estudio revisa críticamente las aplicaciones de la IA en la construcción de pruebas y propone guias prácticas para maximizar sus beneficios y abordar posibles riesgos. **Método:** Se realizó una revisión exhaustiva de la literatura para examinar los avances en aplicaciones basadas en IA en la construcción de pruebas, con énfasis en el desarrollo y calibración de ítems, y se incluyeron ejemplos del mundo real para mostrar su implementación práctica. **Resultados:** Las mejores prácticas para el uso de IA en el desarrollo de pruebas están en evolución, pero requieren supervisión humana. Para generar ítems se necesitan datos de calidad, alineación con el uso previsto, comparación de modelos y validación. Para calibrar, hay que definir el constructo, optimizar las instrucciones (prompts), verificar la alineación semántica, realizar análisis factoriales pseudoexploratorios y evaluar el ajuste del modelo. **Conclusiones:** Se propone una guía práctica que vincula los desafíos de validez, fiabilidad y equidad con recomendaciones para una implementación responsable y eficaz.

Cite as: Suárez-Álvarez, J., He, Q., Guenole, N., & D´Urso, D. (2026). Using artificial intelligence in test construction: A practical guide. *Psicothema, 38*(1), 1-12. https://doi.org/10.70478/psicothema.2026.38.01

Corresponding author: Javier Suárez-Álvarez, suarezj@umass.edu

Artificial Intelligence (AI) is being adopted globally at an unprecedented pace. ChatGPT alone reached 800 million weekly users by April 2025, achieving 90% of its current global user base in just three years. In comparison, the Internet took over 23 years to reach the same level of global adoption (Meeker et al., 2025). Most importantly, its capabilities are still evolving. The Organisation for Economic Co-operation and Development (OECD, 2025) established an independent committee of experts, which estimated that it has reached only about half of its full potential (OECD, 2025). As AI continues to grow, finding ways to use it effectively while reducing potential risks is a major focus for governments, researchers, and practitioners. Educational and psychological assessments are no exception as AI is transforming how tests are designed, delivered, and interpreted.

Educational and psychological assessments are crucial for both individual and societal progress, as they support the identification of needs and the monitoring of progress over time. However, as emphasized in the *Standards for Educational and Psychological Testing* jointly developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), assessments must be relevant, valid, and fair to be effective (AERA, APA, & NCME, 2014). Historically, the improvement of these assessments has progressed alongside advances in methodology and technology. For example, in the 20th century, standardized testing provided a systematic method for evaluating the skills and knowledge of large populations (Sireci et al., 2025). Optical scanners later automated the scoring process, enhancing efficiency and reducing errors. Computer-adaptive testing (CAT) advanced the measurement field by adjusting test difficulty based on individual performance, optimizing the accuracy and relevance of assessments for each test-taker (Zenisky & Sireci, 2002).

Traditional test development followed a rigorous process that typically began with defining the assessment purpose and construct to be measured, manually crafting assessment items, and refining them based on pilot studies and psychometric analysis (AERA et al., 2014; Downing & Haladyna, 2006; Lane et al., 2016; Muñiz & Fonseca-Pedrero, 2019). While this systematic approach is still considered the gold standard for creating relevant, valid, fair measurement tools, it does have its drawbacks. Crafting assessment items manually is time-consuming and often expensive, particularly when done by experienced subject-matter experts (SMEs). Additionally, if the assessments' purpose and construct are innovative and groundbreaking such as AI literacy or prompt engineering, finding the appropriate SMEs can be challenging, which limits accessibility for the broader research community (European Commission, OECD, & Code.org., 2025). Another common challenge is generating a sufficiently large pool of items from which to create parallel versions of tests to counteract item content becoming public online (Bißantz et al., 2024). Designing assessments that reflect test takers' funds of knowledge and cultural backgrounds to enhance engagement, and performance is particularly challenging in traditionally developed assessments, due to rigid blueprints, administration conditions, and high development costs (Walker et al., 2023). Traditional test development is also at an increasing risk of assessing skills that humans routinely use machines to perform (Swiecki et al., 2022).

To address these limitations, researchers have long proposed the use of Automated Item Generation (AIG) and predicting item parameters based on item attributes. AIG enables the creation of diverse item versions based on item templates, reducing item reuse and improving cost efficiency (Bejar et al., 2002; Luecht, 2025). Similarly, statistical modeling approaches have been recommended for decades to estimate item complexity by assigning a difficulty score based on item attributes, allowing developers to systematically predict item performance without relying on extensive field testing (Embretson, 1983, 1999; Sheehan & Mislevy, 1994; Sheehan et al., 2006). These analytical methods offer the potential to streamline development by replacing large-scale pilot studies with model-based predictions. However, it is only with recent technological advancements in generative and representational AI using embeddings that these approaches are beginning to realize their full operational potential (see Table 1 for key operational definitions).

In recent years, the automation of test content generation has significantly streamlined the traditionally manual and costly development processes (Attali et al., 2022; Gierl & Haladyna, 2012; von Davier et al., 2024). Automated scoring systems are now routinely used for evaluating constructed responses - a task that previously required human judgment (von Davier et al., 2022; Yamamoto et al., 2019). When well-design prompts are used, large language models (LLM) can enhance efficiency and quality over traditional automated item generation methods (Bezirhan & von Davier, 2023). LLMs can also be used to obtain item parameters estimates prior to collecting empirical data (Feng et al., 2025; Guenole et al., 2024, 2025). AI technologies are helping to define and refine new constructs, like AI literacy, computational thinking, and prompt engineering, that are becoming increasingly important in digital learning environments (European Commission, OECD, & Code.org., 2025). The use of AI enables the development of innovative item formats such as interactive simulations, scenario-based assessments, and chat-based dialogues (Foster & Piacentini, 2023). AI algorithms can be used to map assessment items to learning standards or curriculum frameworks, thereby assisting with instructional alignment and reducing the burden on subject-matter experts (Butterfuss & Doran, 2025). AI supports adaptive testing and personalized learning paths that respond to individual learner characteristics (Arslan et al., 2024; Sireci et al., 2024; Suárez-Álvarez

Table 1
Key Definitions of AI-Driven Methods in Educational and Psychological Assessment

Name	Description	Example
Generative AI (GenAI)	A class of AI models that can generate new content, such as text, images, or code, based on learned patterns from data.	ChatGPT (OpenAI, 2023)
Machine Learning (ML)	A subset of AI that enables systems to learn from data and improve performance on tasks without being explicitly programmed.	Neural Networks (von Davier, 2018).
Natural Language Processing (NLP)	A field of AI focused on enabling machines to understand, interpret, and respond to human language.	Analyzing students' written responses to assess problem-solving strategies (Yaneva von & Davier, 2023).
Large Language Model (LLM)	A type of NLP model trained on massive text to generate and understand human-like language.	GPT-4 or Claude 3 Opus (OpenAI, 2023; Anthropic, 2024)

et al., 2024; Yan et al., 2024). Digital assessments also capture log (process) data, providing invaluable insights into test takers' cognitive processes and engagement with tasks (He et al., 2021, 2023; Ulitzsch et al., 2023; Suárez-Álvarez et., 2022). Although log (process) data has primarily been used to refine estimates of test takers' proficiencies (Pohl et al., 2021; Wise et al., 2021), it can also be employed to identify item attributes and predict item performance.

The goal of this paper is to summarize current best practices in the applications of Generative AI in modern educational and psychological test construction, specifically focusing on item generation and item calibration. These applications are emphasized because they offer significant benefits in terms of cost efficiency and scalability within educational and psychological assessments, and they also present potential threats to reliability, validity, and fairness. Although these applications have been predominantly utilized by large corporations like Duolingo (von Davier et al., 2024), their adoption among the wider research and practitioner community remains limited. The mission of this paper is to disseminate the latest technological advancements to a broader audience, ensuring that these innovations benefit a diverse group and contribute to the development of a wide range of groundbreaking assessments. Finally, a cautionary commentary is included, outlining strategies to maximize the benefits of AI-driven methods in test construction while minimizing potential risks.

Generative AI in Educational Assessment

Generative AI (GenAI hereafter) has emerged as an innovative tool rapidly adopted across various professional fields, efficiently managing repetitive and time-consuming tasks. Education assessment has been significantly transformed by these advancements, with GenAI becoming a contemporary trend in education. AI facilitates interactive and authentic assessment formats, including simulations, virtual reality (VR) integration, and gamified learning experiences. Automated grading and instant feedback reduce teachers' workloads while enabling personalized learning experiences (Mao et al., 2024). Educational chatbots, also known as educational conversational agents (ECAs), are designed to assist teachers, enhance students' learning processes, and evaluate their performance (Chang et al., 2023). Some chatbots are student-oriented, serving as personalized learning assistants that guide students to answers, evaluate their responses, and foster engagement (Kuhail et al., 2023). Others are tailored to support teachers by preparing class materials, managing course schedules, and tracking deadlines (Ramandanis et al., 2023). The applications of GenAI are widely utilized across various subjects, adapting to different educational formats and needs. In this section we describe emerging methods in educational assessments that leverage GenAI for Automated Item Generation (AIG) and summarize current best practices for implementing them.

Automated Item Generation (AIG)

Automated item generation (AIG) has long been a subject of study in employment and educational assessments (Bejar et al., 2002). Creating test questions—especially for medical licensing and certification—requires significant time and financial resources because it depends on expert input for writing scenarios and crafting credible answer choices. Technologies like machine learning or AI that could help lower these development costs are of great interest to test creators. Traditionally, AIG has focused either on non-verbal formats like visual matrix puzzles (Embretson, 1999), or on techniques resembling fill-in-the-blank exercises similar to MadLibs. Since then, GenAI has significantly transformed both reading and language assessment.

In Maas's (2024) recent research, the team applied a fine-tuned Conditional Transformer Language (CTRL) model to generate English reading comprehension questions for educational purposes, with a focus on controllability and alignment to classroom needs. The model was trained on the Reading Comprehension dataset from Examinations (RACE) and clustered latent traits to allow educators to specify desired question types, for example, cloze-style, title-related, or general questions. The training helped improve the generation of questions tailored to specific reasoning skills. The research found that while the fine-tuned model demonstrated promising results in generating relevant and contextual reading questions, challenges such as overfitting and maintaining consistency in generated outputs remain. This required further refinement for practical classroom adoption (Maas, 2024). Another study compared human-designed and AI-generated English reading comprehension materials, using tools like Twee and Kimi to generate multiple-choice questions based on middle school materials. This research used mixed methods by using both quantitative data and qualitative data to explore the human-AI collaboration in comprehension questions generation. The results of the study showed that the AI tool was significantly more time-efficient, requiring only a fraction of the time needed by the human teacher to complete the task, while generating material of comparable quality, although the human was superior in terms of clarity, relevance, and consistency of the questions with the educational objectives. The study also proved that AI tools can effectively complement teachers in content creation, enhancing efficiency while requiring human guidance to ensure pedagogical depth and appropriateness for classroom contexts (Jen et al., 2024).

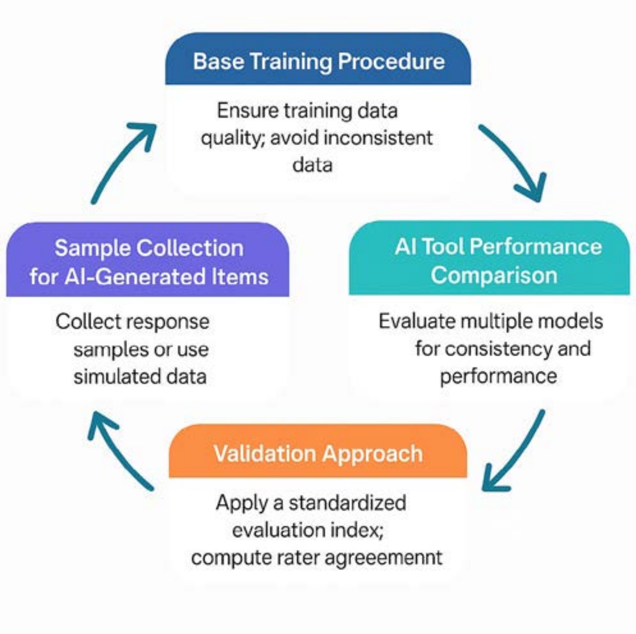
In addition to the Generative Pre-trained Transformer (GPT) model, widely used for text generation through applications like ChatGPT, the BERT model, which underlies Google's search engine capabilities, has also been widely discussed. For example, Kumar's study combined GPT and BERT in a two-stage architecture to improve the coherence and contextual accuracy of automated text generation. Before training, the team preselected models from GPT, Large Scale Decision-Making (LSDM), and Gated Recurrent Units (GRU) and finally selected GPT as the text generation model. After fine-tuning the model with metrics like Bilingual Evaluation Understudy (BLEU) Score and perplexity to gauge the model's performance, the combined model outperformed the single model across various tasks like question-answering and summarization. The research indicated the potential of combining several models for better AI-driven content creation for future diverse applications (Kumar et al., 2024). GenAI chatbots were also powerful tools for language learning and adaptive questions generation during the learning process. Yang et al. (2022) implemented Ellie, a task-based AI voice chatbot, to support Korean EFL students in practicing English speaking. The chatbot fostered meaningful conversations and achieved high task success rates, with students positively perceiving it as a fun and effective learning tool despite some technical and comprehension challenges. The results highlight the potential of AI chatbots to enhance language education while recommending further development to address usability issues and expand application scope (Yang et al., 2022). Von Davier (2018) used a recurrent neural network (RNN) trained on 3,000 test items from the International Personality Item Pool (IPIP) database (Goldberg, 1999), which shows the initial framework of modern test design with a collaboration between human and AI.

Earlier studies noticed that due to limitations in models and data, practical AI-driven AIG was still far off, though the models have been well developed with machine learning techniques. However, as previously noted, the field advanced rapidly when researchers replaced recurrent networks with self-attention-based architectures (Vaswani et al., 2017), enabling simpler designs that support parallel training and allow models to be pre-trained on broad text data before being adapted to specific tasks.

Real-World Example: NAEP Reading Passage Generation

To illustrate how GenAI can support item development, we present an example from the U.S. National Assessment of Educational Progress (NAEP) focused on the generation of reading passages. This process includes ensuring high-quality and consistent training data, evaluating multiple AI models for performance and reliability, applying standardized validation metrics, and collecting response samples to test and refine newly generated items (Figure 1).

Figure 1
Cyclical Framework for Generative AI-Based Test Development


Base Training Procedure
Ensure training data quality; avoid inconsistent data
AI Tool Performance Comparison
Evaluate multiple models for consistency and performance
Validation Approach
Apply a standardized evaluation index; compute rater agreeemennt
Sample Collection for AI-Generated Items
Collect response samples or use simulated data

A recent analysis of NAEP reading tasks revealed inconsistencies in readability scores across the training data. We curated reading passages from NAEP-released items spanning Grades 4 and 8, covering the years 1992 to 2020. To maintain consistency in item design, we focused exclusively on text-based passages paired with multiple-choice questions, deliberately excluding content that incorporated tables or figures. This process yielded 24 passages for Grade 4 and 23 passages for Grade 8. To assess the difficulty level and establish a robust base sample, we applied four widely accepted readability indices: Average Reading Level Consensus, Automated Readability Index (Smith & Senter, 1967), Flesch-Kincaid Grade Level (Kincaid et al., 1975), and SMOG Index (McLaughlin, 1969). Contrary to expectations, the results revealed minimal distinction between grades—approximately 75% of the passages exhibited similar readability scores, making them indistinguishable in terms of grade-level appropriateness.

Inconsistencies such as these can introduce substantial variability in model performance. Moreover, training on biased or misaligned data risks reinforcing and amplifying those biases in model outputs. This is especially concerning when employing general-purpose pre-trained models, where human oversight becomes essential to ensure cultural relevance, fairness, and appropriateness.

To address these challenges and construct a clearly defined, representative training set, we collaborated closely with item developers. Together, we identified and selected six prototypical passages for each grade to serve as the foundation for model training. Figure S1 (Supplementary Material) shows the results from four readability metrics before and after the selection process. It apparently shows a smaller variance after the careful selection for training data. This more accurate training set significantly contributes to the accuracy of AI generation results. It is noted that AI generated results kept at the comparable level as the training set index results. The Fleisch Kincaid Grade Level index consistently showed the lowest value of readability compared with their peers.

NAEP reading passage generation findings indicate that AI-generated nonfiction passages demonstrate a significantly higher difficulty level than fiction passages. This discrepancy likely stems from the inherent variability and creative divergence of fiction writing, which contrasts with the more structured nature of nonfiction texts. Figure S2 (Supplementary Material) presents AI-generated fiction and nonfiction passages for Grade 4. While the nonfiction passages exhibit relatively higher readability scores across all indices—suggesting a level above Grade 4—the fiction passages more closely match the required difficulty range.

To improve the performance of AI in generating fiction content, augmenting the input prompts has shown promise. For example, including explicit labels such as "fiction" or "nonfiction" during training, and emphasizing genre-specific textual features in the prompts, can help guide the AI towards producing passages more consistent with training expectations. These refinements contribute to marginal improvements in readability scores and better alignment with task design.

In this example, we trained AI models using LLMs implemented in ChatGPT, Meta AI, and Claude to generate 40 new passages for Grade 4 and Grade 8 respectively. The readability of these AI-generated passages was reassessed to determine whether they matched the target grade levels. To enhance generation quality, we employed an iterative approach to prompt engineering. Initially, we provided a general description of key differences between Grade 4 and Grade 8 reading levels, including vocabulary complexity, sentence structure, and word count. Our preliminary prompts led to AI-generated passages that mimicked these linguistic features but did not consistently align with expected readability index score ranges. To refine the process, we revised our prompts by explicitly quantifying readability standards, detailing the significance of readability indices, and explaining how they are calculated. This structured approach improved alignment with actual readability levels. Among the three AI tools, ChatGPT demonstrated the most effective performance in passage generation, particularly when utilizing customized GPT functions. The language in the reading passage generated from ChatGPT shows richer descriptions and is highly consistent with the grade level indexes.

As pointed out earlier, we used the consistent evaluation method by using the four readability indicators. This evaluation standard is unchanged between human and AI generated items. As Figure S2 (c) shows (Supplementary Material), the language in the reading passage generated from ChatGPT shows richer descriptions and is highly consistent with the grade level indexes.

Finally, we invited human item developers to help validate the generated items by giving multiple dimensions and calculated the consistency. Though there was no real data collected to validate the items, the experienced human developers give a relatively objective evaluation. In the future study, it is highly recommended to consider using simulated data and/or new sample data collection to make a further validation on the passages.

Practical Guide for Generative AI-Based Test Development

This section provides a practical guide (Table 2) for developing tests using GenAI, aimed at maximizing relevance, validity, and fairness throughout the test construction process.

1. Ensure Consistency and Quality in Training Data

Ensuring the quality of the training dataset is essential for conveying accurate information during the learning process. All materials must undergo rigorous review to confirm the inclusion of high-quality items before they are used for AI training (AERA et al., 2014; Downing & Haladyna, 2006; Lane, Raymond, & Haladyna, 2016; Muñiz & Fonseca-Pedrero, 2019). This step is vital to support critical learning and clear representation of labels in the model.

2. Align AI Use with Intended Uses and Task Type

When using AI for item generation, it is essential to consider both the intended use and the nature of the task. AI models tend to excel at rule-based or logic-driven tasks, yet they often struggle with fiction and emotionally nuanced content. Tasks that require complex human emotion or creativity typically demand additional validation to ensure quality and appropriateness.

3. Compare Multiple AI Models for Reliability

To ensure consistent and reliable outcomes, it is highly recommended to employ at least two AI models and carefully evaluate their performance. Comparing outputs, such as those from ChatGPT and the Claude model, can help identify discrepancies, assess robustness, and improve the overall quality of generated items.

4. Apply a Standardized Validation Approach

Use a consistent evaluation index to assess both training and AI-generated outputs. This ensures alignment with baseline standards and allows for meaningful performance comparisons. Treat AI-generated responses as those from a "human" rater to calculate inter-rater agreement. For example, by verifying whether passages fall within the same readability grade level. This guideline aligns with and extends general guidance on evidence for test validation (Sireci & Benítez, 2023) specifically to AI-based assessments.

5. Verify and Validate AI-Generated Items

While collecting new human response data to evaluate freshly generated items is the most rigorous validation approach, it may not always be feasible due to cost and time restriction. In AI contexts, "verification" often denotes confirming that AI systems are working correctly internally before submitting them for validation scrutiny. This involves checking that AI algorithms generate items as intended, free from technical errors, bias, or unintended patterns, which creates an additional layer addressing the "black box" nature of AI compared to traditional assessment development. For example, consider using AI-simulated data to calibrate item parameters and compare them with the training set (e.g., through Differential Item Functioning analysis), or apply NLP techniques to measure semantic distance between AI-generated items and the original dataset to ensure content alignment and diversity.

Generative AI in Psychological Assessment

GenAI is increasingly applied in psychological assessment and practice, with examples ranging from enhancing diagnostic accuracy and therapeutic interventions in clinical psychology (De la Fuente & Armayones, 2025) to using ChatGPT as a simulated patient to support interactive training and skill development (Sanz et al., 2025). Recent advances in Representational AI using embeddings and GenAI have led to novel approaches in psychological assessment, offering alternatives to traditional self-report methods and enhancing item development, and validation. Generative models (decoders) help create text, such as test items, while representational models (encoders) convert text into numerical formats (embeddings) for analysis. This approach offers a promising way to modernize and improve measurement in psychology (Wulff & Mata, 2025). These embeddings can be used in methods like Pseudo Factor Analysis (PFA) to explore psychological constructs and address issues such as overlap between scales (Guenole et al., 2025). On the other hand, Large Language Models (LLMs) such as GPT-4o and Claude 3 can be used to predict correlations between personality items more accurately than human experts (Schoenegger et al., 2025). Another application comes from Fan et al. (2023), who examined the psychometric properties of personality scores inferred by AI chatbots. These scores, derived from users' free-text input during conversational interactions, showed acceptable reliability and convergent validity but limited discriminant and criterion-related validity. Yuan et al. (2024) examined how users perceive personality scores generated by AI chatbots compared to traditional self-report questionnaires. While users found both methods similarly satisfying and accurate, they tended to view the survey-based results as more trustworthy, likely due to their greater familiarity and simplicity. Sun et al. (2024) presented a framework for developing and validating an AI chatbot based on the Big Five personality model. They emphasize the chatbot's ability to elicit rich, narrative responses aligned with psychological constructs and report improved validity outcomes compared to existing tools. In this section we describe emerging methods in psychological assessment that leverage LLMs for scale construction. We discuss item generation, how to check semantic item alignment, and PFA.

Item Generation, Semantic Item Alignment, and Pseudo Factor Analysis (PFA)

When designing a new assessment, conceptual clarification of how the construct is similar to and different from related constructs is an important step. This can occur qualitatively using subject matter experts before data are collected, but LLMs present the possibility to approach this task analytically with sentence encoders. A sentence encoder is a transformer-based model trained on text to produce highly dense numerical representations of sentences in vector form. These representations are commonly known as embeddings. Association measures such as cosine similarity can be used to compare the similarities of embeddings created from construct definitions. This allows practitioners to determine the constructs' semantic positions in a nomological network, in turn allowing us to move to item generation.

One of the most important requirements is designing effective instructions for the AI, known as prompt engineering, to ensure the output aligns with your goals while minimizing hallucinations and misinterpretations. Prompt engineering with few constraints on instructions leads to direct item generation, where we instruct the LLM to generate items measuring the focal construct without restrictions. We can also use guided item generation methods, where we provide detailed instructions about item requirements, such as construct definitions, item templates, and other constraints necessary such as item polarity (Ferrando et al., 2025). Whether direct or guided item generation is used, we can provide or omit example items in the LLM prompt. If no item examples are given, the approach is zero-shot prompting, giving less control over the items that are created. If we do give examples, we refer to the method as few-shot prompting, which grounds the model in the task context.

Despite giving instructions regarding item features, generated items might not always match our criteria. Quality checks can be implemented as constraints during the item generation process itself. Alternatively, items might be checked with a prompting approach post generation. If the number of items is small (e.g. several hundred or fewer) it is feasible to check these manually and ultimately all items should be human reviewed. As suggested in the educational assessment section, LLMs can also be used to check semantic item alignment with construct definitions. To check semantic item alignment, encodings are generated between the items and the construct definitions, and the cosine similarities are calculated. Items should have high similarities with their parent constructs and low similarities with non-parent constructs. High and low here do not have fixed values, item parent similarities and item non-parent similarities need to be interpreted relative to one another.

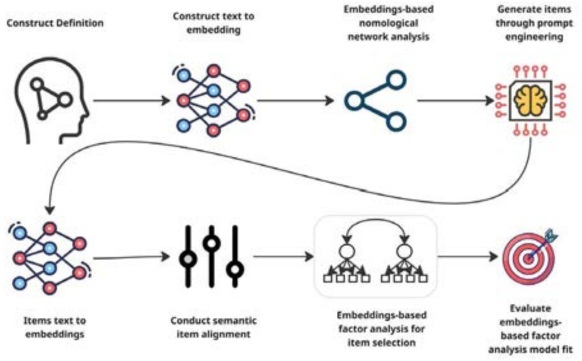
With items generated and pre-screened via semantic item analysis, the factor structure of the items can be examined before responses data are collected with PFA. Similar to traditional factor analysis, PFA allows for different degrees of prior expectations through the use of target rotation. This flexibility enables both fully exploratory analyses, with no prior assumptions, and semi-confirmatory approaches to examine how items group and cluster. At the heart of PFA is the "substitutability assumption", or the idea that the embedding vector for an item statement can stand in for an empirical response vector. This involves forming a cosine similarity matrix between the item embeddings from the previous step, and factor analyzing the matrix in essentially the same way that a correlation matrix of real item responses is analyzed.

Real-World Example: Moral Foundations Scale Calibration

As in the previous section, we use a real-world example to illustrate how GenAI can support AI-based item calibration. This section focuses on the design of a measure targeting executive moral foundations (Graham et al., 2009). Moral foundations are important for senior executives because they make decisions that affect many workers, and these decisions are frequently evaluated in moral terms. Moral foundations are conceptually distinct from familiar industrial psychology constructs, yet they are infrequently included in executive assessment processes. We propose a new moral foundations scale using AI. We show that when our proposed pipeline is followed (Figure 2), PFA can be an effective data-less method for obtaining item pre-knowledge in scale development. We also discuss the challenges relevant to PFA including assessing model fit without sample sizes using raw residuals. We begin our analysis pipeline after we have generated items. More details on the item generation process itself are available in Guenole (2025).

Figure 2

Analytical Pipeline for Generative AI-Based Item Calibration


Construct Definition
Construct text to embedding
Embeddings-based nomological network analysis
Generate items through prompt engineering
Items text to embeddings
Conduct semantic item alignment
Embeddings-based factor analysis for item selection
Evaluate embeddings-based factor analysis model fit

To prepare items for analysis, we first prepared a file of our moral foundations' items (Supplementary Material: factor.csv). We used the MiniLM sentence encoder to generate embeddings of these items in a Jupyter notebook (matrix_generation.ipynb). The notebook uses MiniLM to convert each item into a numerical representation called an embedding, which captures the semantic meaning of the item. Each embedding has hundreds of numbers (dimensions), and the notebook organizes these into columns (one column per dimension). The notebook calculates how similar each item embedding row is to every other item, creating a similarity matrix, much like how you'd calculate correlations between item responses. The output matrix (matrix.csv) can then be prepared for factor analysis by setting any diagonals that are less than one due to rounding errors to 1, as they are in a correlation matrix (matrix.csv). Early theorizing about why this approach works rests on a substitutability assumption (Guenole et al. 2025). This is the notion that a numerical item embedding can substitute for an empirical item response vector under certain conditions.

Next, a factor analysis can be performed on the similarity matrix in R (pfa.R) using any extraction and rotation method. Maximum likelihood estimation with oblique rotation, which allows the resulting factors to be related to each other, have been shown to work well in earlier work. The output includes familiar results from traditional factor analysis, such as eigenvalues, a scree plot, and a pattern loading matrix showing which items load onto which factors. While we present the factor analysis for the final item set, we intentionally included about twice as many items as we intended to keep. This gave us the flexibility to run several rounds of analysis, removing items that didn't load well on any factor or that cross-loaded on multiple factors. After each round of removal, we updated the matrix and repeated the analysis to refine the item set. The items, embedding code, and R code to produce the final factor model are included in Supplemental Materials.

Most methods conventionally used to decide on item retention in the context of EFA can be used with PFA. In the current example we soon discuss, we proposed ensuring that items have their highest loading on their parent factor; that this loading is higher than its loading on any other factor; that this loading is higher than its average loading across all other factors; and that its loading is higher than the average of all other item loadings on that factor. From the pattern matrix in Table S1 (Supplementary Materials) we see that this is the case for most items of the newly developed executive moral foundation scale. From the scree plot in Figure S3 (Supplementary Materials), we see that six factors are plausible, which in fact was the expectation at the outset.

One important point about this approach is that the factor analysis is based on the embedding similarities rather than human responses and therefore there is no sample size. Sample sizes are required for many model-based fit tests and indexes. It is not recommended to simply assume an arbitrarily large sample size, because model fit statistics are influenced by sample size and the correct sample size is required. In this case, we recommend using model free and exploratory approaches to checking model fit based on interpreting the raw residuals. There are several exploratory approaches that might be useful depending on the goal and we describe these here now.

We first plot a heat map of the residual correlations. What we hope to see is that most residual correlations are white indicating they are near zero. We do not want to see any obvious patterns with blocks of blue or red indicating systematically low or high residual correlations between the items after conditioning on the latent factors. In Figure S4 (Supplementary Materials) we see this is mostly the case. We might also plot the distribution of off-diagonal elements of the residual correlation matrix, expecting to see relatively small residuals with few outliers. Again, this appears mostly the case in Figure S5 (Supplementary Materials). Finally, we may choose to plot the original versus the residual correlations. Ideally, we would see a horizontal band of residuals clustered around zero, which is broadly what we see in Figure S5 (Supplementary Materials). We also calculated the Root Mean Square Residual (.037) and the Common Part Accounted for (CAF, Lorenzo-Seva et al., 2011) (.87) which are both indicative of good fit.

Critically, we do not yet present empirical relations with actual factor loadings from participant responses, and this is always an important step. Earlier work by Guenole et al. (2025) shows that pseudo factor loadings are related to empirical loadings, but this is an important next step for the executive moral foundations assessment. We also note while the pseudo and empirical loadings themselves have been shown to be highly correlated. The pseudo factor loadings do not yet differentiate reverse keyed items in the way conventional items do, because cosine similarities between embeddings tend to be positive. Nonetheless, it is still critical to compare pseudo factor structures derived from embeddings with empirical factor structures based on human responses. Ultimately, the empirical factor structure remains the gold standard. Once empirical data are available, alignment between models can be assessed using quantitative metrics such as Tucker's congruence coefficient (values > .85 indicate fair similarity; > .95 indicate strong alignment) and correlation coefficients between corresponding factors (Guenole et al., 2025). Readers may also wish to explore alternative approaches to assessing item dimensionality and discrimination through embedding-based network models (Russell-Lasalandra et al., 2024).

Practical Guide for Generative AI-Based Item Calibration

This section provides a practical guide (Table 2) for item calibration using GenAI, aimed at maximizing relevance, validity, and fairness throughout the test construction process.

6. Use Sentence Encoders to Establish Semantic Construct Validity

Before item generation, clarify how the target construct is similar to or distinct from related constructs. By comparing the semantic similarity of construct definitions within a nomological network, developers can validate construct boundaries early in the design process, improving alignment and focus on subsequent item development.

7. Apply Prompt Engineering Strategies for LLM-Based Item Generation

When generating non-cognitive assessment items with LLMs, use prompt engineering strategies that match the desired level of control. Guided prompts with examples (few-shot) offer greater precision, while minimal prompts without examples (zero-shot) allow more creativity but less control. The choice should reflect the specificity and psychometric standards required for the assessment.

8. Conduct Semantic Item Alignment to Ensure Construct Relevance

To ensure AI-generated items align with the intended construct, apply semantic alignment checks either during or after item generation. This can involve manual review or LLM-based methods, such as calculating cosine similarity between item and construct embeddings. Items should show relatively higher similarity to their target construct than to unrelated ones, guiding item selection and refinement.

9. Use Embedding-Based Factor Analysis with Iterative Refinement for Item Selection

To evaluate AI-generated items, convert item text into embeddings using an LLM and analyze the resulting similarity matrix with factor analysis. Begin with a large item pool to allow for iterative refinement, removing items with weak or cross-loadings. Assign items to factors using systematic criteria based on loading strength and distinctiveness. Ensure the process is transparent and reproducible using shared data and code.

10. Use Model-Free Exploratory Techniques to Evaluate Fit in Embedding-Based Factor Analysis

When factor analyzing item embeddings without response data, traditional fit indices can't be used due to the lack of a sample size. Instead, apply model-free exploratory methods such as heatmaps of residual correlations, distributions of off-diagonal residuals, and plots comparing original to residual correlations to assess whether the latent structure fits the data well.

Table 2
Practical Guide to Generative AI–Based Test Development and Calibration

Generative AI-Based Application	Guidelines
Test Development	*1. Ensure Consistency and Quality in Training Data*
	2. Align AI Use with Intended Uses and Task Type
	3. Compare Multiple AI Models for Reliability
	4. Apply a Standardized Validation Approach
	5. Verify and Validate AI-Generated Items
Item Calibration	*6. Use Sentence Encoders to Establish Semantic Construct Validity*
	7. Apply Prompt Engineering Strategies for LLM-Based Item Generation
	8. Conduct Semantic Item Alignment to Ensure Construct Relevance
	9. Use Embedding-Based Factor Analysis with Iterative Refinement for Item Selection
	10. Use Model-Free Exploratory Techniques to Evaluate Fit in Embedding-Based Factor Analysis

Maximizing Benefits While Reducing Risks

As public trust and engagement in standardized testing declines (Borgonovi & Suárez-Álvarez, 2025; Suárez-Álvarez et al., 2024), AI-driven methods, such ML, NLP, and LLM (see Table 1 for definitions), are being increasingly applied to optimize traditional measurement approaches (Hao et al, 2024; Yaneva & von Davier, 2023). While these innovations offer important gains in efficiency, cost, and scalability, there is a risk that, without also addressing broader concerns of trust, equity, and relevance, educational and psychological measurement may become increasingly disconnected from evolving scientific standards, societal needs, and ethical principles (Burstein et al., 2025; Johnson et al., 2025; Walker et al., 2023). Therefore, to fully harness the benefits of technological innovations like AI in promoting individual and societal progress, it is essential to understand their limitations (Bulut et al., 2024; Dixon-Roman, 2024; Dumas, Greiff, & Wetzel, 2025; Hao et al., 2024; Ho, 2024; Yan, Greiff et al., 2024; Swiecki et al., 2022).

The following section summarizes current limitations of AI-based methods for test construction, organized into four key areas: validity (explainability), reliability (consistency, and generalizability), fairness (training data quality), and data security and privacy. Each issue is linked to specific guidelines to support implementation. However, given the conceptual and practical overlap among these issues and the guidelines to address them, some level of interaction between them is to be expected.

Validity and the "Black Box" Problem

One of the most pressing validity concerns is the lack of transparency in how large AI models make predictions, a challenge often referred to as the *black box problem*. Unlike theory-driven methods grounded in Karl Popper's falsifiability principle, where a scientific theory must be testable and subject to empirical disconfirmation, data-driven AI models do not typically allow for such scrutiny. While these models can serve valuable roles in educational and psychological measurement, the absence of a clear theoretical foundation increases the risk of speculative or spurious conclusions. Rather than discarding theory when confronted with data inconsistencies, we argue for refining theoretical frameworks using advanced methodologies. Empirical inquiry should be guided, and at minimum verified, by theory, not divorced from it.

Furthermore, Explainable Artificial Intelligence (XAI) aims to make AI models more transparent and interpretable, addressing concerns related to model opacity and validity (Samek et al., 2017). By providing clear and understandable explanations of how decisions are made, XAI helps build trust and facilitates validation, particularly in high-stakes domains. This approach has shown promising results in healthcare, improving both clinician understanding and patient outcomes (Doshi-Velez & Kim, 2017; Holzinger et al., 2019). Given these successes, there is growing interest in applying XAI techniques to the educational (Khosravi et al., 2022) and psychological fields (Joyce et al., 2023) to enhance the interpretability and acceptance of AI-driven assessment tools. Our current efforts focus on adapting XAI methods to support transparent and valid test development processes.

Guideline 4 directly addresses the validity concern by establishing systematic methods for evaluating whether AI-generated outputs align with intended constructs. It helps make the AI's decision-making process more interpretable and transparent, reducing the "black box" nature of the model. *Guideline 5* supports construct validity by ensuring that the generated items are actually measuring what they are intended to measure. Through expert review, semantic alignment, or empirical validation, this step helps mitigate the opacity of the model's outputs. *Guideline 6* helps clarify how constructs are defined and differentiated prior to item generation, enhancing conceptual transparency. *Guideline 8* ensures that generated items align with the intended construct, providing a data-driven check on construct representation. Finally, *Guideline 9* offers a framework for analyzing the dimensionality of AI-generated items, thereby supporting construct validity through empirical evidence.

Reliability and the "Hallucination" Problem

Another major threat is (un)reliability. AI models can produce errors, respond inconsistently to identical prompts, and struggle with abstract reasoning, logical inference, or unfamiliar content, issues commonly referred to as *hallucinations*. Although *Guidelines 2* and *3* are intended to mitigate these risks by encouraging task-model alignment and multi-model comparisons, consistent human verification remains essential (see also *Guidelines 4* and *5*).

Guideline 7 recommends using prompt engineering strategies that align with the intended purpose to structure, and guide prompts effectively. This approach reduces variability, increases the consistency of AI-generated items, and is also expected to enhance validity. *Guideline 9* advises applying embedding-based factor analysis iteratively to identify and remove items with weak or inconsistent loadings, thereby enhancing item stability and internal consistency. Finally, *Guideline 10* encourages the use of model-free exploratory techniques to empirically assess internal consistency and dimensional coherence. These methods help identify unreliable or poorly fitting items and support improvements to both internal consistency and the underlying structure of the scale.

Fairness and the "Alignment Gap"

Fairness is compromised when pre-trained models, such as those behind ChatGPT, are used without scrutiny of the cultural responsiveness of their training data. This *alignment gap* reflects a disconnect between model training and intended test use. When sufficient task-specific data are available, *Guideline 1* recommends training models directly on curated, high-quality content. However, when relying on general-purpose pre-trained models, extreme caution is warranted. Human oversight and review are essential to ensure cultural relevance and appropriateness (see *Guidelines 4 and 5*). Our approach maintains a clear boundary between AI-based assessments and the ultimate decision-making responsibilities of psychologists and educators, reinforcing that AI serves as an aid rather than a substitute.

Guideline 6 also aims to ensure that constructs are clearly defined and culturally grounded, helping to reduce the risk of biased construct representation. *Guideline 8* recommends systematically evaluating whether items accurately reflect the target construct across diverse populations. Additionally, *Guideline 7* supports greater control over content generation by incorporating constraints that promote inclusivity and cultural responsiveness.

Data Security and Privacy

Although not directly related to validity, reliability, and fairness, data privacy and security are crucial ethical considerations. Consumer-facing tools like ChatGPT may use submitted prompts and generated responses to further train their models. This poses risks when test content or sensitive data are entered into such platforms. Also, the legal and ethical aspects of content ownership generated by AI warrant future discussion to inform policy and practice.

This issue is addressed through strong data governance practices that ensure sensitive information used in AI-assisted test construction is protected throughout the development process. This includes establishing clear protocols for data access, ensuring compliance with privacy regulations, avoiding the use of open-access consumer AI tools that may reuse input data (such as ChatGPT's free version), and using secure environments for storing and processing both training data and AI-generated content. Effective governance also involves transparency in how data are handled and ensuring that personal or confidential educational data are not inadvertently exposed or misused.

Concluding Remarks

GenAI holds great promise for transforming assessments by enabling faster, more adaptive, and scalable test development. Techniques like embedding-based item evaluation can streamline early test design and reduce costs, helping bridge the gap between semantic AI models and traditional psychometric practices (Guenole et al., 2025; Russell-Lasalandra et al., 2024). However, these innovations must be implemented with caution. Risks such as academic misconduct, technical vulnerabilities, and disciplinary skepticism highlight the need for thoughtful integration (Alasadi et al., 2023; Dolenc et al., 2024; Farrelly et al., 2023; Wang et al., 2023). Crucially, the effectiveness of AI-based tools depends on their alignment with core psychometric principles. Without clear evidence of reliability, validity, and fairness, even the most advanced systems remain superficial. Moving forward, assessment professionals must balance innovation with rigorous empirical standards and ethical safeguards to ensure responsible use of GenAI.

Authors Contributions

Javier Suárez-Álvarez: Conceptualization, Writing - Original draft. **Qiwei He**: Conceptualization, Writing - Original draft. **Nigel Guenole**: Conceptualization, Software, Writing - Original draft. **Damiano D'Urso**: Software, Writing - Review & editing

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Declaration of Interests

The author(s) declare(s) that there is no conflict of interest

Data Availability Statement

Supplementary material for this article is available online in the following link: https://osf.io/fvzyx/?view_only=27238879597b42f984ec7e7b2c721041

References

Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education, 100*(8), 2965–2971. https://doi.org/10.1021/acs.jchemed.3c00323

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

Anthropic. (2024). Claude 3 Opus [Large language model]. https://www.anthropic.com

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence, 7,* 1460651. https://doi.org/10.3389/frai.2024.1460651

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5,* 903077. https://doi.org/10.3389/frai.2022.903077

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). A feasibility study of on-the-fly item generation in adaptive testing. *ETS Research Report Series*, i-44.

Bezırhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5*, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bißantz, S., Frick, S., Melinscak, F., Iliescu, D., & Wetzel, E. (2024). The potential of machine learning methods in psychological assessment and test construction. *European Journal of Psychological Assessment, 40*(1), 1–4. https://doi.org/10.1027/1015-5759/a000817

Borgonovi, F. & Suárez-Álvarez, J (2025). *How can adult skills assessments best meet the demands of the 21st century?*. OECD Social, Employment and Migration Working Papers, No. 319. OECD Publishing. https://doi.org/10.1787/853db37b-en

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation, 5*(3), 3. https://doi.org/10.59863/MIQL7785

Burstein, J. (2025, April 17). *The Duolingo English Test responsible AI standards (Duolingo Research Report No. DRR-25-05)*. Duolingo. https://englishtest.duolingo.com/research

Butterfuss, R., & Doran, H. (2025). An application of text embeddings to support alignment of educational content standards. *Educational Measurement: Issues and Practice, 44*(1), 73–83. https://doi.org/10.1111/emip.12581

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability, 15*(17), 12921.

De la Fuente, D., & Armayones, M. (2025). AI in psychological practice: What tools are available and how can they help in clinical psychology? *Psychologist Papers, 46*(1), 18-24. https://doi.org/10.70478/pap.psicol.2025.46.03

Dixon-Román, E. (2024). AI and psychometrics: Epistemology, process, and politics. *Journal of Educational and Behavioral Statistics, 49*(5), 709–714. https://doi.org/10.3102/10769986241280623

Dolenc, K., & Brumen, M. (2024). Exploring social and computer science students' perceptions of AI integration in (foreign) language instruction. *Computers and Education: Artificial Intelligence, 7*, 100285.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv.* https://doi.org/10.48550/arXiv.1702.08608

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Erlbaum Associates Publishers.

Dumas, D., Greiff, S., & Wetzel, E. (2025). Ten guidelines for scoring psychological assessments using artificial intelligence [Editorial]. *European Journal of Psychological Assessment, 41*(3), 169–173. https://doi.org/10.1027/1015-5759/a000904

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433.

European Commission, OECD, & Code.org. (2025, May). *Empowering learners for the age of AI: An AI literacy framework for primary and secondary education* (Review draft). https://www.oecd.org/digital/empowering-learners-ai-literacy-framework

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., … Hack, E. (2023, January 5). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *PsyArXiv.* https://doi.org/10.31234/osf.io/pk2b7

Farrelly, T., & Baker, N. (2023). Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences, 13*(11), 1109.

Feng, W., Tran, P., Sireci, S., & Lan, A. S. (2025). *Reasoning and sampling-augmented MCQ difficulty prediction via LLMs*. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education. AIED 2025* (Lecture Notes in Computer Science, Vol. 15880). Springer, Cham. https://doi.org/10.1007/978-3-031-98459-4_3

Ferrando, P. J., Morales-Vives, F., Casas, J. M., & Muñiz, J. (2025). Likert scales: A practical guide to their design, construction and use. *Psicothema, 37*(4), 1–15. https://doi.org/10.70478/psicothema.2025.37.24

Foster, N. & Piacentini, M (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation*. Routledge. https://doi.org/10.4324/9780203803912

Goldberg, L. R. (1999). *A broad-bandwidth, public domain personality inventory measuring the lower-level facets of several five-factor models*. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), Personality Psychology in Europe (Vol. 7, pp. 7-28). Tilburg University Press

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029.

Guenole, N., D'Urso, E. D., Samo, A., Sun, T., & Haslbeck, J. (2025).Enhancing scale development: Pseudo factor analysis of language embedding similarity matrices. *PsyArXiv*. https://osf.io/preprints/psyarxiv/vf3se_v2

Guenole, N., Samo, A., Sun, T. (2024). Pseudo-Discrimination Parameters from Language Embeddings. *OSF*. https://osf.io/9a4qx_v1

Guenole, N. (2025). *Psychometrics.ai: Transforming Behavioral Science with Machine Learning*. https://psychometrics.ai

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice, 43*(2), 16-29. https://doi.org/10.1111/emip.12602

He, Q., Borgonovi, F., Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers and Education, 166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2023). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning, 39*, 719–736. https://doi.org/10.1111/jcal.12748

Ho, A. D. (2024). Artificial intelligence and educational measurement: Opportunities and threats. *Journal of Educational and Behavioral Statistics, 49*(5), 715-722. https://doi.org/10.3102/10769986241248771

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312. https://doi.org/10.1002/widm.1312

Jen, F.-L., Huang, X., Liu, X., & Jiao, J. (2024). *Can generative AI really empower teachers' professional practices? Comparative study on human-tailored and GenAI-designed reading comprehension learning materials*. In L. K. Lee, P. Poulova, K. T. Chui, M. Černá, F. L. Wang, & S. K. S. Cheung (Eds.), *Technology in Education. Digital and Intelligent Education. ICTE 2024. Communications in Computer and Information Science, vol. 2330* (pp. 112–123). Springer.

Johnson, M. S. (2025, April). *Responsible AI for measurement and learning: Principles and practices* (ETS Research Report No. RR-25-03). ETS Research Institute.

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digital Medicine, 6*(6). https://doi.org/10.1038/s41746-023-00751-9

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* (Research Report No. 56). Institute for Simulation and Training. https://stars.library.ucf.edu/istlibrary/56

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*, 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Kumar, P., Manikandan, S., & Kishore, R. (2024). *AI-driven text generation: A novel GPT-based approach for automated content creation*. 2024 2nd International Conference on Networking and Communications (ICNWC). IEEE.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development (2nd ed.)*. Routledge.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

Luecht, R. M. (2025). *Assessment engineering in test design: Methods and applications* (1st ed.). Routledge. https://doi.org/10.4324/9781003449464

Maas, A. C. (2024). *An empirical study on training generative AI to create appropriate questions for English reading comprehension* [Doctoral dissertation, Tohoku University]. Tohoku University Repository.

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. TechTrends, *68*(1), 58-66.

Meeker, M., Simons, J., Chae, D., & Krey, A. (2025). *Trends – artificial intelligence (AI).* BOND. https://www.bondcap.com/report/tai/

McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading, 12*(8), 639-646.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema, 31*(1), 7–16. https://doi.org/10.7334/psicothema2018.291

OECD (2025). *Introducing the OECD AI capability indicators.* OECD Publishing. https://doi.org/10.1787/be745f04-en

OpenAI. (2023). GPT-4 technical report. https://arxiv.org/abs/2303.08774

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340. https://doi.org/10.1126/science.abd3300

Ramandanis, D., & Xinogalos, S. (2023). Designing a chatbot for contemporary education: A systematic literature review. *Information, 14*(9), 503.

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2024, September 12). Generative psychometrics via AI-GENIE: Automatic item generation and validation via network-integrated evaluation. *PsyArXiv.* https://doi.org/10.31234/osf.io/fgbj4

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv.* https://doi.org/10.48550/arXiv.1708.08296

Sanz, A., Tapia, J. L., García-Carpintero, E., Rocabado, J. F., & Pedrajas, L. M. (2025). ChatGPT simulated patient: Use in clinical training in Psychology. *Psicothema, 37*(3), 23-32. https://doi.org/10.70478/psicothema.2025.37.21

Schoenegger, P., Greenberg, S., Grishin, A., Lewis, J., & Caviola, L. (2025). AI can outperform humans in predicting correlations between personality items. *Communications Psychology, 3*, 23. https://doi.org/10.1038/s44203-025-00123-1

Sheehan, K. M., Kostin, I., & Persky, H. (2006, April). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP Grade 8 Reading Assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA. Educational Testing Service.

Sheehan, K., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills (ETS RR-94-14).* Educational Testing Service.

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217-26. https://doi.org/10.7334/psicothema2022.477

Sireci, S. G., Crespo Cruz, E, Suárez-Álvarez, J, & Rodríguez Matos, G. (2025). *Understanding UNDERSTANDardization research.* In R. Bennett, R., L. Darling-Hammond & A. Barinarayan (Eds.), Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy, Routledge. https://doi.org/10.4324/9781003435105

Sireci, S. G., Suárez-Álvarez, J., Zenisky, A. L., & Oliveri, M. E. (2024). Evolving educational testing to meet students' needs: Design-in-real-time assessment. *Educational Measurement: Issues and Practice, 43*(4), 112–118. https://doi.org/10.1111/emip.12653

Smith, E. A., & Senter, R. J. (1967). *Automated readability index* (Vol. 66, No. 220). Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Suárez-Álvarez, J., Fernández-Alonso, R., García-Crespo, F. J., & Muñiz, J. (2022). The use of new technologies in educational assessments: Reading in a digital world. *Psychologist Papers, 43*(1), 36–47. https://doi.org/10.23923/pap.psicol.2986

Suárez-Ávarez, J., Oliveri, M. E., Zenisky, A., & Sireci, S. G (2024). Five key actions for redesigning adult skills assessments from learners, employees, and educators. *Journal for Research on Adult Education, 47*, 321–343. https://doi.org/10.1007/s40955-024-00288-8

Sun, T, B., Drasgow, F., & Zhou, M. X. (2024, May 1). Developmente and validation of an artificial chatbot to assess personality. *PsyArXiv*. https//doi.org/10.131234/osf.io/ahtr9

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2023). Accounting for careless and insufficient effort responding in large-scale survey data—Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods, 56*(2), 804–825. https://doi.org/10.3758/s13428-022-02053-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://arxiv.org/abs/1706.03762

von Davier, A. A., Runge, A., Park, Y., Attali, Y., Church, J., & LaFlair, G. (2024). The item factory: Intelligent automation in support of test development at scale. In H. Jiao & R. W. Lissitz (Eds.), *Machine learning, natural language processing, and psychometrics* (Marces Book Series) (pp. 1– 25). Information Age Publishing Inc.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

von Davier, M., Tyack, L., & Khorramdel, L. (2022). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement, 83*(3), 556–585. https://doi.org/10.1177/00131644221098021

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles* (ETS RR-23-11). Educational Testing Service. https://doi.org/10.1002/ets2.12374

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *Open Journal of Computer Science, 4*, 280–286. https://doi.org/10.48550/arXiv.2305.18339

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment, 26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, 1-11. https://doi.org/10.1038/s41562-024-02089-y

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Introduction of multistage adaptive testing design in PISA 2018.* OECD Education Working Papers, No. 209, OECD Publishing, https://doi.org/10.1787/b9435d4b-en

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*, 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

Yaneva, V., & von Davier, M. (Eds.). (2023). *Advancing natural language processing in educational assessment* (1st ed.). Routledge. https://doi.org/10.4324/9781003278658

Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL, 34*(3), 327–343. https://doi.org/10.1017/S0958344022000039

Yuan, L. (I.), Sun, T., Dennis, A. R., & Zhou, M. (2024). Perception is reality? Understanding user perceptions of chatbot-inferred versus self-reported personality traits. *Computers in Human Behavior: Artificial Humans, 2*, 100057. https://doi.org/10.1016/j.chbah.2024.100057

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362. https://doi.org/10.1207/S15324818AME1504_02