EAM2025 XI Conference
23RD - 25TH JULY 2025
European Association of Methodology
Spain Tenerife Canary Islands

Innovating Test Development with Generative AI:A Content Validity Study in AIL

Xavier G. Ordóñez (UCM) & Sonia J. Romero Martínez (UNED)

Introduction

The accelerated emergence of generative artificial intelligence (AI) has opened new avenues in educational measurement. Specifically, large language models (LLMs) enable unprecedented efficiency in generating and evaluating test items. However, few studies have examined whether AI can serve as both the creator and evaluator of items within a rigorous psychometric framework. This study addresses that gap by introducing a fully automated protocol for item development and validation, using generative AI exclusively. The protocol was applied to the domain of AI literacy, which includes understanding, applying, and ethically evaluating AI in everyday life (Ng et al., 2021; Kong et al., 2024). The study aligns with the revised Bloom's taxonomy to ensure cognitive diversity and examines content validity through classic indices (CVR, CVC, Aiken's V). It seeks to evaluate the feasibility of AI-led test construction, while acknowledging the importance of human judgment in interpretive and ethical domains (Messick, 1995).

Objectives

Methodological:

- Design a systematic AI-driven item development protocol
- Evaluate agreement among AI models
- Assess content validity using CVR, CVC, Aiken's V

Applied:

- Create a robust AI literacy item bank
- Refine items based on clarity, alignment, and usefulness

Method

Phase 1: Item Generation

- Items were constructed based on Bloom's revised taxonomy (Anderson & Krathwohl, 2001) and AI literacy frameworks by Ng et al. (2021), Kong et al. (2024), and Chiu et al. (2023).
- Six cognitive levels were considered: Remember, Understand, Apply, Analyze, Evaluate, Create.
- Items targeted six application domains: general AI, virtual assistants, translation, social media, entertainment, and document editing.
- A total of 720 items were generated using ChatGPT with standardized prompts and controlled parameters (temperature = 0.7).

Phase 2: Item Evaluation

- A panel of 8 generative AI models (Gemini, Claude, LLaMA, Mistral, etc.) rated each item.
- Step 1: Congruence analysis — each model assigned a Bloom level to each item. Items with full agreement (8/8) were retained.
- Step 2: Psychometric rating — items were rated on clarity, alignment, wording, discrimination, and usefulness using a 5-point Likert scale.
- Indices used: CVR (Tristán-López, 2008), CVC (Hernández-Nieto, 2002), and Aiken's V (Aiken, 1980).

Results

1. Congruence Analysis
Each item was independently classified by eight AI models according to the most appropriate level of Bloom's revised taxonomy. Items were retained if there was complete agreement (8/8 models), resulting in 203 items with high dimensional alignment. Dimensions such as Evaluate and Understand exhibited the highest agreement across domains like social media and entertainment.
2. Content Validity Indices
The retained items (203 from Bloom-based dimensions and 180 from ethics/awareness) were evaluated across five attributes: clarity, theoretical alignment, discrimination, wording, and usefulness. Content validity was quantified using three indices:

- **CVR** (Content Validity Ratio): threshold of 0.58 based on Tristán-López (2008).
- **CVC** (Content Validity Coefficient): values above 0.80 indicated strong clarity and relevance.
- **Aiken's V**: values above 0.70 deemed acceptable (Davis, 1992), with 96% of Evaluate items surpassing 0.80.

A total of 323 items met the validity thresholds. Highest scores were obtained in the Evaluate and Create dimensions, while Remember and Apply yielded weaker psychometric performance.

Conclusions

Generative AI enables scalable and efficient test item development. However, human judgment remains essential for interpretive and ethical decisions in educational measurement.
This study demonstrates that large language models can serve not only as item generators but also as evaluators, replicating key steps traditionally assigned to expert panels. The high content validity observed in items aligned with higher-order cognitive processes (Evaluate and Create) reinforces the potential of AI-assisted design for competency-based assessment.
Despite these promising results, some dimensions such as Remember and Apply showed limitations in discriminative power, suggesting that AI-generated items may be less effective for lower cognitive levels. Additionally, while inter-model agreement was strong in several domains, variability still emerged in ethically and socially complex areas—underscoring the irreplaceable role of human expertise in final test validation.
Overall, the proposed protocol offers a replicable framework for future research and practical applications in educational measurement. It encourages a hybrid approach where generative AI streamlines large-scale item production, but human oversight ensures interpretive rigor and contextual sensitivity.

References

- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Chiu, T. K., et al. (2023). Systematic literature review on AI in education. *Computers and Education: Artificial Intelligence*, 4, 100118.
- Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5(4), 194–197.
- Hernández-Nieto, R. A. (2002). Contributions to statistical analysis. Universidad de Los Andes.
- Kong, S. C., et al. (2024). Artificial Intelligence literacy in education. *Innovations in Education and Teaching International*.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.
- Ng, D. T. K., et al. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041.
- Tristán-López, A. (2008). Modificación al modelo de Lawshe. *Avances en Medición*, 6, 37–48.

LaTeX TikZposter

Universidad de La Laguna
Gobierno de Canarias
Consejería de Universidades, Ciencia e Innovación y Cultura
Agencia Canaria de Investigación, Innovación y Sociedad de la Información
Cabildo de Tenerife
Tagoro
Instituto Canario de Igualdad
tea hogrefe cajasiete