REVIEW OF CRIME, PEACE AND SOCIETY
ISSN (O): 3079-6105, ISSN (P): 3079-6091
Vol. 3, No. 4, 2026, 55-60

ITEM DISCRIMINATION AND RELIABILITY OF HUMAN-DESIGNED VERSUS AI-GENERATED ACHIEVEMENT TESTS: A CLASSICAL TEST THEORY COMPARISON

Abdul Azeem[1],Prof. Dr. Muhammad Naseer Ud Din[2], Dr. Abdul Wahab[3], Shakeel Nawaz[4]

Abstract

The rapid adoption of large language models for educational content creation has outpaced the empirical evidence needed to judge whether AI-generated multiple-choice items meet established psychometric standards, particularly at the graduate level and within under-represented higher-education systems. Using Classical Test Theory, this study addressed a single, focused research question: do human-designed and Claude AI-generated multiple-choice achievement tests, matched for content and item difficulty, differ in item discrimination and internal-consistency reliability? Two 30-item parallel forms were constructed from an identical Table of Specifications aligned with Bloom's Revised Taxonomy and administered, in a within-subjects design, to 200 graduate students at Kohat University of Science and Technology, Pakistan. Item discrimination was estimated using the upper–lower index and corrected point-biserial correlation; internal consistency was estimated using KR-20. Item difficulty was compared first, as a precondition check, before testing the primary hypothesis. The two forms did not differ significantly in mean item difficulty (human M = .555, AI M = .544; t (29) = 1.324, p = .196), confirming a fair basis for comparison. The AI-generated form showed significantly higher mean discrimination (M = .531 vs. .476; t (29) = −2.384, p = .024) and significantly higher internal-consistency reliability (KR-20 = .850 vs. .806; 95% CI for the difference .015, .076). Under matched content and difficulty conditions, Claude AI-generated items discriminated more effectively between higher- and lower-performing students and produced a more internally consistent achievement test than human-designed items. These findings offer localized, quantitative evidence that generative AI can meet or exceed core Classical Test Theory benchmarks for graduate-level achievement testing, supporting cautious, expert-supervised integration of AI-assisted item development in resource-constrained higher-education contexts.

Article History
Received 15 Jue 2026
Revised 25 June 2026
Accepted 14 July 2026
Published 20 July 2026

OPEN ACCESS

Corresponding Author *

Keywords
artificial intelligence; automated item generation; item discrimination; test reliability; Classical Test Theory; KR-20; graduate assessment

Introduction

Multiple-choice questions (MCQs) remain the dominant format for large-scale achievement testing because they are efficient to administer, objective to score, and adaptable across cognitive levels when carefully constructed (Haladyna & Rodriguez, 2013). Constructing psychometrically sound MCQs, however, is labor-intensive, requiring subject expertise, iterative item review, and empirical piloting (Ali & Zahra, 2024). The emergence of large language models capable of drafting assessment items from a structured prompt has attracted considerable interest in reducing this

[1] M.Phil. Scholar Institute of Education and Research, Kohat University of Science and Technology azeemscholar86@gmail.com
[2], Director, Institute of Education and Research, Kohat University of Science and Technology. dr.naseeruddin@kust.edu.pk
[3], Lecturer, Institute of Education and Research, Kohat University of Science and Technology, abdulwahab@kust.edu.pk
[4], PhD scholar/Visiting Lecturer, Institute of Education and Research, Kohat University of Science and Technology shakeelnawaz@kust.edu.pk

burden (Kurdi et al., 2020). Whether AI-generated items meet the psychometric standards required of graduate-level assessment, however, remains an open and empirically contested question.
Evidence to date is mixed. Some studies report that AI-generated items approximate human-authored items in difficulty and discrimination once mapped to an explicit test blueprint (Kiyani et al., 2025; Linde et al., 2026), while others report that AI-generated items cluster within a narrower difficulty range or discriminate less effectively than expert-written items (Laupichler et al., 2024; Küchemann et al., 2024). Most of this evidence originates from Western or East Asian, and predominantly medical-education, settings (Kıyak & Emekli, 2024); comparatively little is known about how these tools perform within Pakistani public-sector higher education (Khurshid et al., 2024; Habib et al., 2024).

This study addresses that gap by comparing a human-designed and a Claude-AI-generated achievement test built for the same graduate-level research-methods course at Kohat University of Science and Technology (KUST). Rather than attempting to evaluate every dimension of assessment quality simultaneously, the study is deliberately scoped around a single, focused research question: ***RQ:*** *Do human-designed and Claude AI-generated multiple-choice achievement tests, matched for content and item difficulty, differ in item discrimination and internal-consistency reliability?*

Answering this question under Classical Test Theory (CTT) provides direct, actionable evidence for institutions weighing whether AI-assisted item generation can meet the core statistical benchmarks expected of operational achievement tests.

Literature Review

Classical Test Theory situates an examinee's observed score as the sum of a true score and measurement error, and evaluates test quality primarily through item difficulty, item discrimination, and internal-consistency reliability (Crocker & Algina, 2008). Discrimination indices commonly the upper lower index or the point-biserial correlation indicate how effectively an item separates higher- from lower-performing examinees and are considered a primary criterion for item retention (Ebel & Frisbie, 1991; Kim et al., 2024). Internal-consistency reliability is most estimated via Cronbach's coefficient alpha for continuous or Likert-type data (Cronbach, 1951) and its dichotomous special case, the Kuder–Richardson Formula 20, for right/wrong scored items (Kuder & Richardson, 1937).

A growing body of research has applied these indices to AI-generated MCQs. Kiyani et al. (2025) benchmarked ChatGPT-generated items against faculty-authored items in dental education and found broadly comparable psychometric performance once items were reviewed against a blueprint. Similarly, Linde et al. (2026) reported that GPT-4o-generated items achieved discrimination and difficulty profiles comparable to human-authored items across imaging specialties. Law et al. (2025), in a cohort study embedded within a high-stakes medical examination, likewise found AI-generated items performed competitively with human-generated items on classical indices. In contrast, Laupichler et al. (2024) found that ChatGPT-generated exam questions differed measurably from human-generated items in academic medicine, and Küchemann et al. (2024) reported reliability and validity concerns for ChatGPT-generated concept-inventory items. Kıyak and Emekli (2024), reviewing prompting strategies for MCQ generation, concluded that psychometric performance depends heavily on prompt structure and blueprint alignment rather than being an inherent property of the underlying model.

Reviews of automatic question generation more broadly (Kurdi et al., 2020) and of large language models for examination generation (Artsi et al., 2024) converge on a similar conclusion:

AI-generated items can achieve strong statistical performance, but this performance is contingent on structured prompting and downstream empirical verification rather than guaranteed by the technology itself. This body of evidence remains concentrated in medical and Western or East Asian educational contexts; within Pakistani higher education specifically, evidence remains limited despite growing AI adoption (Khurshid et al., 2024; Habib et al., 2024; Oad et al., 2024). The present study contributes direct, CTT-based evidence on discrimination and reliability from a Pakistani public-sector graduate programme, addressing this specific gap.

Methodology

Research Design

A quantitative, within-subjects comparative design was used. Two 30-item parallel MCQ forms one human-designed, one generated by Claude AI were constructed from an identical Table of Specifications aligned with Bloom's Revised Taxonomy, ensuring that any difference in psychometric performance could be attributed to the item-generation method rather than to differences in content coverage or cognitive-level distribution (Haladyna & Rodriguez, 2013).

Participants

Two hundred graduate students enrolled in the course Understanding Research Methods at the Institute of Education and Research, KUST, were selected via simple random sampling. All participants completed both test forms, enabling paired comparisons on every item.

Instruments

Both forms comprised 30 dichotomously scored MCQs (1 = correct, 0 = incorrect) covering eight instructional units. A pilot study involving 30 additional graduate students informed minor item revisions prior to main administration, following standard item-analysis decision criteria for retention, revision, or replacement.

Procedure

Institutional approval and informed consent were obtained prior to data collection. Both forms were administered under standardized examination conditions in a single sitting.

Data Analysis

Item difficulty (p) was calculated as the proportion of correct responses per item and compared between forms using a paired-samples t-test, to confirm that any subsequent discrimination or reliability difference was not confounded by a pre-existing difficulty gap. Item discrimination was estimated using the upper lower 27% discrimination index and the corrected item-total (point-biserial) correlation and compared between forms using a paired-samples t-test on the 30 matched items. Internal-consistency reliability was estimated for each form using KR-20 (Kuder & Richardson, 1937), and the difference between coefficients was evaluated using a 95% confidence interval. All analyses were conducted in IBM SPSS 26 at $\alpha = .05$.

Results

Equivalence of Item Difficulty (Precondition Check)

Mean item difficulty did not differ significantly between the human-designed ($M = .555$) and AI-generated ($M = .544$) forms, $t(29) = 1.324$, $p = .196$. This confirms that the two forms presented students with a statistically equivalent level of challenge, providing a fair basis for comparing discrimination and reliability.

Item Discrimination

The AI-generated form obtained a significantly higher mean discrimination index ($M = .531$) than the human-designed form ($M = .476$), $t(29) = -2.384$, $p = .024$. Twenty-nine of the 30 AI-

generated items were classified as having very good discrimination, compared with 25 of 30 human-designed items; neither form contained a marginal, poor, or negatively discriminating item. Corrected point-biserial correlations converged with this pattern (AI M = .365 vs. human M = .311).

Internal-Consistency Reliability

The AI-generated form obtained a significantly higher KR-20 coefficient (.850) than the human-designed form (.806); the 95% confidence interval for the difference, [.015, .076], excluded zero, indicating a statistically meaningful advantage for the AI-generated form.

Table 1

Index	nan-Designed (Test H)	I-Generated (Test AI)	Statistic	Sig.
lean item difficulty (p)	.555	.544	t (29) = 1.324	p = .196
ean discrimination index (D)	.476	.531	(29) = −2.384	p = .024
ın corrected point-biserial (r_a)	.311	.365		
-20 reliability coefficient	.806	.850	% CI [.015, .076]	

Discussion

The equivalence of item difficulty between the two forms indicates that, when both are constructed from an identical Table of Specifications, human-designed and AI-generated items can present students with a comparable level of challenge consistent with reports that AI-generated items closely approximate human-authored items in difficulty when grounded in an explicit blueprint (Kiyani et al., 2025; Linde et al., 2026). Against this equivalent baseline, the AI-generated form's significantly higher discrimination and reliability provide a cleaner test of the technology's core measurement performance than a comparison confounded by differing difficulty would allow.

These results align with a growing body of evidence that large language models, when appropriately prompted against a structured blueprint, can produce items that function at least as well as, and in some cases better than, human-authored items on classical psychometric indices (Law et al., 2025; Linde et al., 2026). At the same time, other researchers have reported that AI-generated items can discriminate less effectively or cluster within a narrower difficulty band than expert-written items (Laupichler et al., 2024; Küchemann et al., 2024), underscoring that AI performance on these indices is not uniform across contexts, subject domains, or prompting strategies, and that empirical verification within each new setting remains necessary (Kıyak & Emekli, 2024).

The higher reliability obtained by the AI-generated form is consistent with the interpretation that items generated from a single, internally consistent prompting process may cohere more tightly around the intended construct than items authored individually by a single expert across sessions, though this mechanism was not directly tested here and warrants dedicated investigation. Within the broader landscape of AI adoption in Pakistani higher education, where empirical evidence on AI-assisted assessment remains limited (Khurshid et al., 2024; Habib et al., 2024; Oad et al., 2024),

these findings offer localized, quantitative evidence that can inform how institutions such as KUST might responsibly incorporate AI-assisted item generation, consistent with broader calls for structured, blueprint-driven, and empirically verified use of generative tools in assessment development (Kurdi et al., 2020; Artsi et al., 2024).

Conclusion

Under conditions of matched content coverage and equivalent item difficulty, Claude AI-generated multiple-choice items discriminated significantly more effectively between higher- and lower-performing graduate students and produced a significantly more internally consistent achievement test than items designed by a human subject-matter expert. These findings directly answer the study's single research question: yes, the two item-generation methods differ in discrimination and reliability, with the AI-generated form outperforming the human-designed form on both indices. This should not be read as evidence that human-designed items are deficient both forms achieved good-to-very-good discrimination and good reliability by conventional standards but rather as evidence that structured, blueprint-driven AI item generation can meet or exceed core Classical Test Theory benchmarks under the conditions tested here.

Recommendation for Future Research

This study was conducted at a single institution, within a single course, using a single AI model and prompting protocol; the discrimination and reliability advantage observed here may not generalize to other subjects, item-generation prompts, model versions, or student populations. All psychometric indices were computed under Classical Test Theory, which is inherently sample-dependent; replication using Item Response Theory would allow discrimination and difficulty parameters to be estimated with less dependence on the specific sample. Future research should also examine whether the reliability advantage observed for AI-generated items persists after independent human expert revision, and whether it holds across repeated test administrations and diverse subject domains.

References

Ali, K., & Zahra, D. T. (2024). Tips for effective use and quality assurance of multiple-choice questions in knowledge-based assessments. European Journal of Dental Education, 28(3), 655–662. https://doi.org/10.1111/eje.12992

Artsi, Y., Sorin, V., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2024). Large language models for generating medical examinations: Systematic review. BMC Medical Education, 24, 354. https://doi.org/10.1186/s12909-024-05239-y

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice-Hall.

Habib, M. M., Hoodbhoy, Z., & Rehman, A. (2024). Knowledge, attitudes, and perceptions of healthcare students and professionals on the use of artificial intelligence in healthcare in Pakistan. PLOS Digital Health, 3(5), e0000443. https://doi.org/10.1371/journal.pdig.0000443

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. Routledge.

Imtiaz, M., ud Din, N., Nawaz, S., & Wahab, A. (2025). Examining Peer Learning Factors and their Impact on Research Productivity among Graduate Students at Kohat University of Science and Technology. *Research Journal for Social Affairs*, *3*(6), 1769-1778.

Khurshid, S., Khurshid, S., & Toor, H. K. (2024). Learning for an uncertain future: Artificial intelligence a challenge for Pakistani education system in the era of digital transformation. Qualitative Research Journal. Advance online publication. https://doi.org/10.1108/QRJ-02-2024-0045

Kim, Y. H., Kim, B. H., Kim, J., Jung, B., & Bae, S. (2024). Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2023, Korea: A psychometric study. Journal of Educational Evaluation for Health Professions, 21, 40. https://doi.org/10.3352/jeehp.2024.21.40

Kıyak, Y. S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. Postgraduate Medical Journal, 100(1190), 858–865. https://doi.org/10.1093/postmj/qgae065

Kiyani, A., Hanif, F., Muhammad, M., Iqbal, S., Zaib, N., Bashir, U., & Ali, K. (2025). Benchmarking ChatGPT-generated multiple-choice questions against faculty-authored items in dental education. Scientific Reports, 15, 44805. https://doi.org/10.1038/s41598-025-28492-7

Küchemann, S., Rau, M., Schmidt, A., & Kuhn, J. (2024). ChatGPT's quality: Reliability and validity of concept inventory items. Frontiers in Psychology, 15, 1426209. https://doi.org/10.3389/fpsyg.2024.1426209

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160. https://doi.org/10.1007/BF02288391

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Laupichler, M. C., Rother, J. F., Grunwald Kadow, I. C., Ahmadi, S., & Raupach, T. (2024). Large language models in medical education: Comparing ChatGPT- to human-generated exam questions. Academic Medicine, 99(5), 508–512. https://doi.org/10.1097/acm.0000000000005626

Law, A. K. K., Lo, C. W. H., Yeung, C. Y., Yau, W. H., Lam, T. S. K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: A cohort study in a high-stakes examination. BMC Medical Education, 25, 208. https://doi.org/10.1186/s12909-025-06796-6

Linde, P., Fichter, F., Dietlein, M., Sudbrock, F., Afshar, K., Dapper, H., Fokas, E., Hillebrecht, A.-L., Raupach, T., & Laupichler, M. C. (2026). Psychometric properties and detectability of GPT-4o–generated multiple-choice questions compared with human-authored items across imaging specialties. npj Digital Medicine, 9, 15. https://doi.org/10.1038/s41746-025-02313-7

Nawaz, S., Ud Din, M. N., Wahab, A., & Khan, A. (2025). The impact of artificial intelligence (AI) on the analytical ability skills (AAS) of graduate students at Kohat University of Science and Technology (KUST) Kohat. *International Journal of Multicultural Education*, *27*(2), 38–49.

Nawaz, S., ud Din, N., & Wahab, A. (2025). Exploring the impact of AI-ChatGPT on creativity skills of postgraduate students at KUST Kohat. *ASSAJ*, *3*(01), 916-928.

Oad, L., Shah, R., Sewani, R., & Ahmad, N. (2024). Empowerment of artificial intelligence in learning optimisation: Student perceptions in Karachi, Pakistan. International Journal of Educational Sciences, 47(2), 34–44. https://doi.org/10.31901/24566322.2024/47.02.1374