DOI: 10.71317/jgst.2.9(s2).2026.694
Journal of Global Social Transformation
Journal homepage: https://rjsaonline.org/index.php/JGST

An AI-Enhanced Analytic Rubric for Assessing Higher-Order Thinking Skills in Pakistani Higher Education: A Framework Development and Validation Design

Dr. Ajaz Shaheen[1], Manishah Devi[2], Fatima[3]
[1] Assistant Professor & Head of Department, Faculty of Education, Lasbela University of Agriculture, Water and Marine Sciences (LUAWMS), Uthal, Balochistan, Pakistan.
[2] Department of Teacher Education, LUAWMS.
Email: manisha.edu@luawms.edu.pk
[3] Department of Teacher Education, LUAWMS.
Email: fatima.edu@luawms.edu.pk

ARTICLE INFO

Submission:
August 08, 2026
Revised:
August 22, 2026
Accepted:
September 11, 2026
Available Online:
September 26, 2026

Keywords:
Artificial intelligence; generative AI; higher-order thinking; analytic rubric; assessment; human-AI collaboration; higher education; Pakistan.

Corresponding author:
Manishah Devi
Email:
manisha.edu@luawms.edu.pk

ABSTRACT

Artificial intelligence (AI), particularly generative AI, is changing how higher education institutions design learning activities, produce feedback, and judge student work. The change creates an assessment problem: institutions need credible evidence of higher-order thinking (HOTS), while conventional assessment formats can make it difficult to distinguish students' own reasoning from AI-assisted production. This paper develops the conceptual architecture and validation design of an Artificial Intelligence-Enhanced Analytic Rubric (AI-EAR) for assessing HOTS in Pakistani higher education. The proposed rubric is grounded in the revised Bloom taxonomy, constructive alignment, formative assessment, and human-centred AI principles. It defines five assessable dimensions—analysis, evaluation, creation, metacognitive reflection, and ethical-collaborative reasoning—and assigns AI a bounded decision-support role rather than autonomous scoring. The study proposes four sequential phases: literature and framework synthesis; expert consultation and content validation; pilot administration across purposively selected public and private higher education institutions; and psychometric evaluation through internal consistency, inter-rater reliability, and exploratory factor analysis. Three research questions guide the development and validation process. The paper does not claim empirical validity coefficients because full data collection has not yet been completed. Instead, it presents a transparent, testable validation pathway and specifies the evidence required before the instrument can be recommended for wider institutional use. The framework is designed to support assessment that is criterion-referenced, transparent, ethically bounded, and sensitive to the Pakistani higher-education context.

Introduction

Higher education is increasingly expected to demonstrate that students can analyse evidence, evaluate competing claims, create solutions, monitor their own reasoning, and act responsibly in complex situations. These capabilities are not adequately represented by assessment systems that rely predominantly on recall or highly structured responses. The revised Bloom taxonomy distinguishes remembering, understanding, applying, analysing, evaluating, and creating, with analysing, evaluating, and creating commonly treated as higher levels of cognitive performance (Anderson & Krathwohl, 2001). Constructive alignment similarly requires intended learning outcomes, learning activities, and assessment to be deliberately aligned (Biggs, 1996).

The arrival of generative AI has intensified this assessment challenge. Generative AI can support brainstorming, explanation, feedback, drafting, and problem solving, but the same technologies can also make it difficult for teachers to determine which aspects of submitted work represent the learner's own reasoning. Recent evidence is not uniformly negative or positive. A 2025 meta-

analysis of 29 experimental and quasi-experimental studies reported a moderate overall positive effect of generative AI on higher-order thinking, with stronger effects for problem solving and critical thinking than for creativity (Zhao et al., 2025). This finding supports a more differentiated assessment approach rather than treating HOTS as a single undivided construct.

The Pakistani context adds institutional and policy considerations. The Higher Education Commission's current framework on generative AI in higher education adopts a balanced position: it distinguishes responsible use from misuse, recognizes AI literacy as important, and encourages institutions to develop context-appropriate policies rather than relying on blanket prohibition (Higher Education Commission, n.d.). UNESCO's AI Competency Framework for Teachers also emphasizes human agency, accountability, ethics, AI pedagogy, and professional learning, providing a useful international basis for defining the human role in AI-supported assessment (UNESCO, 2024).

Against this background, this paper proposes an AI-Enhanced Analytic Rubric (AI-EAR). The purpose is not to automate teachers' judgments. Instead, AI is positioned as a bounded support layer that can organize evidence, flag possible inconsistencies, and assist with criterion coverage, while the final judgement remains with the educator. The paper therefore treats AI-supported assessment as a human-AI collaborative process rather than an automated marking system.

Problem Statement

Three related problems motivate the proposed framework. First, Pakistani higher education needs assessment approaches capable of producing defensible evidence of complex thinking while remaining feasible in large and diverse classrooms. Second, the increasing availability of generative AI makes conventional product-based assessment less informative about the student's independent reasoning unless assessment tasks explicitly require evidence of process, reflection, judgement, and responsible AI use. Third, AI-supported marking raises questions about transparency, consistency, contextual judgement, privacy, and academic integrity. Existing analytic rubrics may assess particular cognitive skills, but the proposed study addresses the need for a locally examined framework that combines HOTS dimensions with explicit human-AI role boundaries.

Accordingly, the central research problem is: how can an analytic rubric be designed and systematically validated so that it assesses higher-order thinking in Pakistani higher education while using AI as transparent decision support and preserving educator accountability?

Research Questions

Because this is a framework-development and validation-design study rather than a completed empirical effectiveness study, research questions are appropriate and improve the logical coherence of the paper. They should be framed around development and validation, not around results that have not yet been collected.

RQ1: What dimensions and observable performance indicators should constitute an AI-enhanced analytic rubric for assessing HOTS in Pakistani higher education?

RQ2: To what extent do experts judge the proposed AI-EAR dimensions and performance indicators to have adequate content validity and contextual relevance?

RQ3: What evidence of reliability and construct structure is obtained when the AI-EAR is piloted across selected Pakistani higher education institutions?

Objectives

- To develop a theoretically grounded AI-enhanced analytic rubric for assessing HOTS in higher education.
- To define five assessable HOTS-related dimensions and behaviourally anchored proficiency descriptors.
- To establish the content validity of the proposed rubric through expert review and Content Validity Index (CVI) procedures.
- To pilot the rubric in selected public and private higher education institutions and across relevant disciplines.
- To examine internal consistency, inter-rater reliability, and the preliminary factor structure of the instrument.
- To establish explicit ethical and operational boundaries for AI use in assessment.

Literature Review

Higher-Order Thinking and the Revised Bloom Taxonomy

The revised Bloom taxonomy organizes cognitive processes from remembering and understanding through applying, analysing, evaluating, and creating (Anderson & Krathwohl, 2001). For the present framework, analysing, evaluating, and creating provide the core cognitive foundation. However, the AI-EAR extends this foundation by adding metacognitive reflection and ethical-collaborative reasoning. These additions are not presented as additional levels of Bloom's taxonomy; rather, they represent complementary dimensions needed to interpret complex performance in an AI-mediated learning environment.

Constructive Alignment and Assessment for Learning

Constructive alignment argues that intended learning outcomes, learning activities, and assessment should be deliberately connected so that students are asked to demonstrate the performances that instruction intends to develop (Biggs, 1996). Formative assessment further emphasizes evidence and feedback that can improve learning rather than merely produce a final grade (Black & Wiliam, 1998). An analytic rubric can operationalize this alignment by translating broad outcomes into observable criteria and performance descriptors. The AI-EAR therefore treats the rubric not simply as a marking sheet but as an assessment-and-feedback architecture.

Generative AI and HOTS

Evidence concerning generative AI and HOTS is developing rapidly. Zhao et al. (2025), synthesizing 29 experimental and quasi-experimental studies, found a moderate positive overall effect on HOTS but different patterns across problem solving, critical thinking, and creativity. The implication for assessment is important: an instrument should not assume that AI use has the same relationship with every cognitive domain. Assessment should instead examine the evidence of reasoning produced by the learner in each domain.

AI and Assessment in Pakistan

Recent Pakistani evidence indicates substantial engagement with AI in higher education. Danish et al. (2024), using data from 256 teaching staff in public and private higher education institutions, reported high familiarity with AI and substantial perceived usefulness for teaching and academic work. More recent research has examined AI literacy among 733 university students across five major disciplines in Pakistan and reported moderate overall AI literacy, with variation across disciplines (Assessing AI literacy of university students in Pakistan, 2026). Research published in 2026 has also examined teachers' challenges in assessing student performance in the era of generative AI, underscoring concerns about assessment practices in higher education (Kalim et al., 2026). These studies support the relevance of the problem but do not themselves establish the validity of the proposed AI-EAR. That distinction is important: evidence of AI adoption or perceptions should not be presented as evidence that a new rubric is valid.

Human-Centred and Ethical AI

UNESCO's AI Competency Framework for Teachers identifies 15 competencies across five dimensions: human-centred mindset, ethics of AI, AI foundations and applications, AI pedagogy, and AI for professional learning. The framework emphasizes human agency and accountability rather than replacing teachers with AI (UNESCO, 2024). HEC's framework on generative AI in higher education similarly emphasizes responsible use, academic integrity, AI literacy, and institutionally appropriate policies (Higher Education Commission, n.d.). These principles inform the AI-EAR's central design rule: AI may assist the evaluator, but it should not independently determine the final academic judgement.

Content Validity and Instrument Development

Content validity should be established before a new instrument is treated as a measurement tool. Expert ratings can be summarized using the item-level and scale-level Content Validity Index (I-CVI and S-CVI). Polit, Beck, and Owen (2007) provide a widely used methodological discussion of the CVI and its interpretation. The present study therefore places expert review before pilot administration and requires revision or deletion of weak items before psychometric analysis.

Conceptual Framework

The AI-EAR integrates four complementary foundations: (1) the revised Bloom taxonomy for cognitive complexity; (2) constructive alignment for connection among outcomes, learning activities, and assessment; (3) formative assessment for feedback and learner

development; and (4) human-centred AI principles for responsible, transparent, and accountable AI use. The framework operationalizes these foundations through five dimensions.

Dimension	Assessment focus	Theoretical basis	Permitted AI support	Human responsibility
Analysis	Deconstructing problems, claims, evidence and relationships	Revised Bloom taxonomy	Evidence organization, pattern detection, coherence flags	Judge relevance, context and disciplinary meaning
Evaluation	Judging arguments, evidence and solutions against explicit criteria	Bloom; formative assessment	Criterion coverage, consistency flags, scoring prompts	Make final judgement and contextual weighting
Creation	Producing original, feasible and contextually grounded solutions	Bloom	Idea mapping and similarity/originality prompts	Judge creativity, feasibility and disciplinary/cultural fit
Metacognitive reflection	Explaining and monitoring reasoning, strategies and learning decisions	Constructive alignment; formative assessment	Prompted reflection and trend summaries	Interpret reflection and provide mentoring feedback
Ethical-collaborative reasoning	Responsible AI use, evidence attribution, transparency and collaborative judgement	Human-centred AI; UNESCO AI framework	AI-use logs and transparency checks	Judge integrity, responsibility, values and contextual appropriateness

Each dimension should be represented by four proficiency levels: Emerging, Developing, Proficient, and Advanced. The final operational instrument should contain behaviourally anchored descriptors rather than vague adjectives. For example, an Advanced performance in Evaluation should demonstrate explicit criteria, credible evidence, consideration of counter-evidence, and a justified conclusion; simply producing a fluent answer should not qualify as advanced performance.

Proposed AI-EAR Rubric Architecture

The proposed architecture uses criterion-referenced assessment. Each dimension receives a set of observable indicators and four proficiency levels. A four-level scale is preferable at the framework stage because it avoids an indeterminate midpoint and encourages raters to distinguish meaningful differences in evidence. During pilot testing, however, the scoring model should be empirically examined rather than assumed to be optimal.

A critical design feature is evidence of process. Students should be asked, where appropriate, to document their use of AI, identify AI-generated suggestions that they accepted or rejected, verify important claims, and explain how the final reasoning reflects their own judgement. Such requirements shift assessment from detection of AI use toward demonstration of learning and responsible AI use.

The AI component should be modular and technology-agnostic. The rubric should not depend on one commercial AI system because model capabilities, interfaces, pricing, and policies change. AI outputs should therefore be treated as provisional decision support. No final grade should be assigned solely by an AI system.

Methodology

The study will use a sequential exploratory mixed-methods design. The design begins with conceptual and qualitative development and proceeds to quantitative validation. This sequence is appropriate because the study is developing a new assessment framework rather than testing an already established instrument.

Phase 1: Literature and Framework Synthesis

A structured literature search will identify scholarship on HOTS, analytic rubrics, AI-supported assessment, human-AI collaboration, academic integrity, metacognition, and assessment validity. International policy and framework documents from UNESCO and HEC will be used to define ethical and operational boundaries. The output will be an initial construct map and item pool.

Phase 2: Expert Consultation and Content Validation

A multidisciplinary expert panel should include specialists in educational assessment, teacher education, AI in education, educational technology, and relevant disciplinary teaching. Experts will rate each indicator for relevance, clarity, representativeness, and contextual appropriateness. I-CVI and S-CVI will be calculated. Items that fail the predetermined criterion will be revised, merged, or removed. A qualitative record of expert comments will also be retained so that numerical CVI values are not interpreted without substantive review.

For a defensible journal study, the final protocol should report the number and expertise of experts, the rating scale, the exact CVI calculation, the decision rule for retention/revision/removal, and the changes made after review. These details are essential for reproducibility.

Phase 3: Pilot Administration

The revised AI-EAR will be piloted with a purposively selected sample of students and trained raters from public and private higher education institutions. Institutions should represent differences in disciplinary context and digital infrastructure. Suitable disciplines may include education, social sciences, applied sciences, and other fields in which complex performance can be elicited through authentic tasks.

Pilot tasks should be standardized sufficiently to permit comparison but flexible enough to reflect disciplinary variation. At least two trained human raters should independently score a common subset of student work. The AI-support condition should be documented explicitly, including what AI was allowed to do, what information was supplied to the system, and which decisions remained exclusively human.

Phase 4: Psychometric Evaluation

Internal consistency will be estimated for each dimension and the overall scale. Cronbach's alpha may be reported, but it should not be treated as sufficient evidence of validity; where appropriate, additional reliability estimates such as McDonald's omega can strengthen the analysis. Inter-rater reliability should be evaluated using an appropriate intraclass correlation coefficient (ICC), with the exact model specified. If the study compares conventional and AI-supported rating conditions, the comparison should be based on the same performance evidence and trained raters where feasible.

Exploratory factor analysis (EFA) will be used to examine whether the observed item structure is compatible with the five hypothesized dimensions. Before EFA, the study should report sample adequacy, the correlation matrix, KMO, Bartlett's test, extraction method, factor-retention rule, and rotation. If sample size and study design permit, a subsequent confirmatory factor analysis (CFA) on an independent sample would provide stronger evidence than relying on EFA alone.

Validity should be treated as an argument supported by multiple sources of evidence rather than as a single coefficient. The final empirical paper should therefore distinguish content evidence, response-process evidence, internal-structure evidence, inter-rater reliability, and relationships with relevant external measures where such measures are available.

Ethical Procedures

Institutional ethical approval should be obtained before collection of student or faculty data. Participants should provide informed consent. Student identifiers should be minimized, and AI systems should not receive personally identifiable or confidential

assessment data. AI-use logs, where collected, should be handled under the institution's data-protection requirements. The study should also disclose the AI tools and model versions used during data collection because model behaviour can change over time.

Proposed Scoring and Interpretation

A practical prototype could assign 1–4 points to each proficiency level for each dimension. If all five dimensions are equally weighted, the total score would range from 5 to 20. However, equal weighting should be treated as a provisional design decision rather than a validated fact. Discipline-specific implementations may require different weighting, and empirical evidence should inform whether a total score is defensible.

More importantly, the AI-EAR should preserve a profile of performance across dimensions rather than reducing HOTS to one total number. A student may demonstrate strong analysis but weak ethical reasoning, or strong creation but limited metacognitive reflection. Reporting the profile can therefore be more educationally informative than reporting only an aggregate score.

Anticipated Findings and Testable Propositions

Because data collection has not been completed, the following are propositions rather than findings.

- Expert review is expected to support the relevance and clarity of most proposed indicators, while ethical-collaborative reasoning may require additional refinement because it combines assessment, integrity, and AI-use considerations.
- AI-supported decision assistance is expected to improve evidence organization and scoring consistency, but this must be tested empirically rather than assumed.
- The five-dimensional structure is expected to show meaningful relationships among dimensions while retaining sufficient distinctiveness to justify separate reporting.
- Implementation conditions, including digital infrastructure, faculty training, and institutional policy clarity, are expected to influence feasibility.
- The relationship between AI support and HOTS is expected to vary across dimensions; therefore, the study should avoid treating AI as uniformly beneficial or harmful.

Discussion

The proposed AI-EAR addresses an assessment problem rather than attempting to solve AI use through surveillance. A rubric that explicitly values reasoning, evidence, reflection, and ethical transparency can make assessment more informative even when students have access to generative AI. This approach is consistent with human-centred AI principles because it preserves the educator's responsibility for academic judgement.

The framework also responds to the limitations of simple AI-detection approaches. Detection systems may provide signals about generated text, but detecting AI authorship is not equivalent to measuring learning. The AI-EAR instead asks what the student can demonstrate: how a problem was analysed, how evidence was evaluated, how a solution was created, how reasoning was monitored, and how AI assistance was used responsibly. The distinction is particularly important because the goal of assessment is to establish evidence of learning, not merely to classify text as human- or AI-generated.

For Pakistani HEIs, the framework has potential relevance to outcome-based education, faculty development, quality assurance, and assessment policy. However, these implications should remain conditional until empirical validity and reliability evidence is established. The instrument should not be described as standardized, validated, or universally applicable on the basis of the present conceptual paper alone.

Implications

For teachers

Faculty development should focus on rubric literacy, task design, evidence-based feedback, responsible AI use, and calibration among raters. AI should support—not replace—professional judgement.

For students

Assessment should make responsible AI use visible and teach students to verify outputs, disclose meaningful AI assistance, and explain their own reasoning.

For institutions

HEIs should establish clear assessment policies defining permitted, restricted, and prohibited AI uses and should provide secure infrastructure and faculty training.

For quality assurance

AI-supported assessment should be documented through transparent procedures, validation evidence, rater training, and periodic review rather than treated as a black-box technology.

For future research

Independent-sample CFA, measurement invariance across disciplines and institution types, criterion-related evidence, longitudinal studies, and comparisons of human-only and human-AI-supported scoring should be investigated.

Limitations

- The proposed paper does not yet contain empirical reliability or validity coefficients; these must be generated from the planned pilot study.
- Purposive institutional sampling may limit generalizability beyond participating HEIs.
- AI-supported assessment depends partly on infrastructure, model capability, institutional policy, and faculty AI literacy.
- Rubric performance may vary by discipline and task type; a single generic rubric may therefore require discipline-sensitive examples.
- AI model outputs can change over time, so the AI-support protocol should be versioned and documented.
- Self-reported AI-use information may be incomplete; therefore, self-report should not be treated as a perfect record of AI assistance.

Conclusion

This paper presents the AI-Enhanced Analytic Rubric (AI-EAR) as a proposed framework for assessing higher-order thinking in Pakistani higher education under conditions of increasing generative-AI use. Its distinctive contribution is the combination of five assessable dimensions—analysis, evaluation, creation, metacognitive reflection, and ethical-collaborative reasoning—with a bounded human-AI assessment model. AI is positioned as decision support for evidence organization and consistency checking, while final academic judgement remains with the educator.

The paper should be understood as a framework-development and validation-design study, not as a completed validation study. Its claims therefore concern the proposed architecture and the planned evidence-generating process. The next empirical phase must establish content validity, inter-rater reliability, internal structure, and—where possible—relationships with relevant external measures. Only after such evidence is obtained should AI-EAR be described as a validated instrument or recommended for broad institutional adoption. This distinction strengthens the methodological credibility of the study and prevents conceptual innovation from being presented as empirical evidence.

References

1. Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.

2. Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347–364. https://doi.org/10.1007/BF00138871

3. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

4. Danish, M. H., Joiya, S. A., & Ali, Z. (2024). Implications and usefulness of artificial intelligence in higher education institutions: A case of Pakistan. Bulletin of Education and Research, 46(3), 57–75.

5. Higher Education Commission. (n.d.). Framework on use of Generative AI (GenAI) tools in Higher Education Institutes (HEIs). Government of Pakistan.

6. Kalim, U., Irfan, S., Jamil, B., & Kanwar, A. (2026). Teachers' challenges in assessing student performance in the era of generative artificial intelligence. Discover Education, 5, 379. https://doi.org/10.1007/s44217-026-01370-8

7. Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in Nursing & Health, 30(4), 459–467. https://doi.org/10.1002/nur.20199

8. UNESCO. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084

9. Zhao, Y., Yue, Y., Sun, Z., Jiang, Q., & Li, G. (2025). Does generative artificial intelligence improve students' higher-order thinking? A meta-analysis based on 29 experiments and quasi-experiments. Journal of Intelligence, 13(12), 160. https://doi.org/10.3390/jintelligence13120160

10. Assessing AI literacy of university students in Pakistan: A cross-sectional survey. (2026). Digital Library Perspectives, 42(1), 141–158. https://doi.org/10.1108/DLP-06-2025-0094