HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

METHODOLOGY FOR DESIGNING PEDAGOGICAL TEST ITEMS USING ARTIFICIAL INTELLIGENCE

Raisov Shamsiddin Rashid o‘g‘li
Master’s student in Testology and Knowledge Assessment
Department of Master’s Studies
Termez State Pedagogical Institute, Uzbekistan

Ismailov Murodulla Kaxramonovich
Doctor of Philosophy (PhD) in Pedagogical Sciences, Associate Professor
Termez State Pedagogical Institute, Uzbekistan
E-mail: murodullaismailov89@mail.ru
ORCID: 0009-0003-8309-1488

Abstract: This article examines the methodological foundations of designing pedagogical test items using artificial intelligence. The rapid development of generative artificial intelligence and large language models has created new opportunities for automating several stages of educational assessment, including the formulation of learning outcomes, preparation of test specifications, generation of test items and distractors, classification of questions according to cognitive levels, and preliminary quality control. However, artificial intelligence cannot independently guarantee the validity, reliability, fairness, and pedagogical appropriateness of a test. Therefore, the article proposes a human-centred methodology in which artificial intelligence functions as an auxiliary instrument under the continuous supervision of a teacher, subject-matter expert, and testologist. The proposed methodology includes the definition of the assessed construct, development of a test blueprint, preparation of verified source materials, prompt engineering, AI-assisted item generation, expert review, pilot administration, psychometric analysis, revision, and creation of an item bank.

Keywords: artificial intelligence, pedagogical testing, test item design, educational assessment, testology, generative AI, large language models, Bloom’s taxonomy, prompt engineering, validity, reliability, psychometric analysis.

The digital transformation of education has significantly changed traditional approaches to teaching, learning, monitoring, and evaluating educational outcomes. Contemporary educational institutions increasingly use electronic learning platforms, digital textbooks, automated assessment systems, virtual laboratories, adaptive learning technologies, electronic portfolios, learning analytics, and intelligent tutoring systems. These technologies make it possible to organise the educational process more flexibly, provide rapid feedback, analyse individual learning progress, and improve the objectivity of assessment.

One of the most rapidly developing directions of digital education is the application of artificial intelligence in pedagogical assessment. Artificial intelligence systems can process large amounts of instructional information within a short period, identify important concepts, classify educational content, generate questions in different formats, propose answer options, prepare distractors, create scoring rubrics, and produce alternative versions of tests. These functions may reduce the routine workload of teachers and allow them to devote more attention to interpreting assessment results, providing feedback, and supporting learners.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

At the same time, the introduction of artificial intelligence into test development raises a number of methodological, pedagogical, psychometric, ethical, and technological questions. An AI system may produce a grammatically correct and logically organised item that does not measure the intended learning outcome. It may formulate a question with more than one defensible answer, create implausible distractors, introduce information outside the curriculum, incorrectly determine the cognitive level, or reproduce cultural and linguistic bias. Therefore, the external fluency of an AI-generated task should not be treated as evidence of its scientific or psychometric quality.

Pedagogical testing is not limited to writing a collection of questions. It is a systematic process that begins with defining what should be measured and ends with interpreting the test results. A valid test requires correspondence among curriculum content, learning outcomes, assessment tasks, scoring procedures, administration conditions, and the intended use of the results. Artificial intelligence can assist at several stages of this process, but it cannot independently assume responsibility for the quality and consequences of educational assessment.

The need for a systematic methodology becomes particularly important when AI-generated test items are used in summative examinations, certification, admission, selection, or other assessments that may influence a learner's academic future. In such contexts, every item should be supported by content analysis, expert judgement, pilot data, and psychometric evidence. The process should also protect personal information, ensure fairness, and prevent the uncontrolled disclosure of test materials.

The present article expands the methodological principles and stages outlined in the supplied draft and develops them into a continuous academic analysis. The purpose of the study is to substantiate a comprehensive methodology for designing pedagogical test items using artificial intelligence and to determine the conditions under which AI-assisted assessment can be scientifically reliable, pedagogically useful, and ethically responsible.

The objectives of the research are to analyse the educational potential of artificial intelligence in pedagogical testing, explain the principles of AI-assisted item development, develop a staged model for generating and validating test tasks, clarify the role of prompt engineering, identify expert and psychometric quality criteria, describe the risks associated with automated question generation, and formulate practical recommendations for teachers, testologists, and educational institutions.

Pedagogical testing is a scientifically organised process of measuring learners' knowledge, skills, competencies, attitudes, and educational achievements through standardised tasks and procedures. A pedagogical test is therefore not merely a set of questions prepared for a lesson. It is a measurement instrument designed to collect evidence about a clearly defined educational construct.

The concept of a construct refers to the knowledge, ability, competence, or psychological characteristic that the test is intended to measure. In education, a construct may include knowledge of a particular topic, the ability to apply a rule, competence in solving a professional problem, critical thinking, reading comprehension, or the capacity to evaluate a pedagogical situation. If the construct is not clearly defined, the generated test items may measure unrelated or superficial abilities.

For example, a test intended to assess the ability to apply pedagogical principles should not consist only of questions that require the memorisation of definitions. Similarly, a task designed to assess analytical competence should require the learner to compare information, identify relationships, interpret evidence, or make a justified conclusion. The cognitive operation required by the item must correspond to the intended learning outcome.

The quality of a pedagogical test is generally evaluated through such characteristics as validity, reliability, objectivity, fairness, practicality, and security. Validity concerns the degree to which evidence and theory support the intended interpretation and use of test scores. Reliability refers to the consistency of measurement results under comparable conditions. Objectivity concerns the extent to

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

which scoring and interpretation are independent of the assessor's subjective preferences. Fairness requires that the assessment not create irrelevant advantages or disadvantages for particular groups of learners.

These characteristics are interconnected. A test may produce consistent results but still measure the wrong construct. In this case, the test may be reliable but not valid. Similarly, a well-designed item may lose its measurement quality if administration conditions differ substantially among learners. Therefore, test quality depends not only on question wording but also on the entire assessment system.

Artificial intelligence can support this system by accelerating content analysis, generating alternative tasks, checking language, and organising item metadata. However, professional control is required because artificial intelligence operates primarily by identifying patterns in data and generating probable textual responses. It does not possess an independent understanding of curricular priorities, institutional assessment policy, learner characteristics, or the consequences of test results.

The most appropriate model is not complete automation but structured cooperation between artificial intelligence and human experts. Artificial intelligence is effective in generating alternatives, organising information, proposing distractors, preparing draft explanations, and formatting item metadata.

Human specialists are responsible for defining the construct, approving the blueprint, checking scientific accuracy, confirming the cognitive level, evaluating fairness, interpreting psychometric evidence, and making final decisions.

In this model, the AI system performs generative and organisational functions, while the teacher and testologist perform evaluative and responsible functions. Such distribution preserves human agency and reduces the risk of uncritical automation.

The role of the teacher also changes. The teacher becomes not only an item writer but also a designer of assessment architecture, prompt author, quality controller, data interpreter, and ethical decision-maker. This requires new professional competencies.

Testological competence includes knowledge of validity, reliability, assessment purpose, task formats, scoring, and psychometric analysis. Digital competence includes the ability to use AI tools, construct prompts, verify outputs, protect data, and document the process. Subject competence is required to evaluate scientific accuracy and curricular relevance.

The integration of artificial intelligence into pedagogical testing should be understood as a transformation of professional practice rather than a simple technological innovation. AI can accelerate the production of questions, but speed alone is not an indicator of assessment quality.

The effectiveness of AI-assisted testing should be evaluated through several indicators. These include the proportion of generated items accepted after expert review, correspondence with the blueprint, accuracy of cognitive-level classification, difficulty and discrimination indices, distractor performance, reliability, fairness, error frequency, quality of feedback, and transparency of development.

A scientifically defensible test requires more than AI generation. It requires evidence from content analysis, expert judgement, learner response processes, pilot testing, psychometric analysis, and intended score use. The interpretation of test results should be supported by this evidence.

Artificial intelligence is particularly useful in low-stakes formative assessment, where teachers need frequent questions and rapid feedback. In high-stakes contexts, more rigorous control is required because errors may have significant consequences.

Educational institutions should therefore establish different levels of review depending on the purpose of the assessment. Classroom quizzes may require teacher review, whereas certification and

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

admission tests should involve subject experts, testologists, psychometricians, security specialists, and institutional approval.

Artificial intelligence provides substantial opportunities for improving the design of pedagogical test items. It can analyse instructional content, generate multiple task variants, propose plausible distractors, classify questions by cognitive level, prepare explanations, and support the development of item banks.

However, automatically generated questions do not automatically constitute a valid pedagogical test. Assessment quality depends on construct definition, curriculum alignment, test specifications, expert review, pilot administration, psychometric analysis, fairness, security, and appropriate interpretation of results.

The proposed methodology includes nine interconnected stages: defining the assessment purpose and construct, developing a test blueprint, preparing verified source materials, designing an effective prompt, generating alternative items, conducting expert review, organising pilot administration, performing psychometric analysis, and revising items for inclusion in a documented item bank.

The most effective approach is professionally managed human–AI collaboration. Artificial intelligence should generate, organise, and suggest, while teachers, subject experts, and testologists should verify, evaluate, revise, and make final decisions.

When these conditions are fulfilled, artificial intelligence can become a powerful instrument for creating diverse, efficient, scientifically grounded, and learner-oriented pedagogical assessments. Its value lies not in replacing professional judgement but in strengthening the capacity of educators to design better measurement instruments and make more informed decisions about teaching and learning.

References:

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.
2. Anderson, L. W., and Krathwohl, D. R. (Eds.). (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman.
3. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. New York: David McKay Company.
4. Downing, S. M., and Haladyna, T. M. (Eds.). (2006). Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.
5. Haladyna, T. M. (2004). Developing and Validating Multiple-Choice Test Items. Mahwah, NJ: Lawrence Erlbaum Associates.