A STEP-BY-STEP METHODOLOGY FOR DESIGNING AND INTEGRATING TEST SOFTWARE IN INFORMATICS EDUCATION

Madinabonu Sayfiddin qizi Fayzullayeva
is a fourth-year student at Bukhara Pedagogical Institute.

Abstract

The design and integration of test software in informatics instruction represents a complex, multi-disciplinary challenge situated at the intersection of psychometrics, didactics, and computer science. This paper presents a systematic, evidence-based methodology for developing and deploying educational test programs in secondary and higher informatics education. Drawing on Classical Test Theory (CTT), Item Response Theory (IRT), and Bloom's Revised Taxonomy, the study establishes a seven-stage ADDIE-aligned development model — encompassing needs analysis, design, item authoring, platform selection, implementation, piloting, and iterative improvement — and maps it onto a principled item classification framework. The methodology was applied in the context of Uzbekistan's national informatics curriculum, producing a validated item bank across four cognitive levels and three difficulty tiers. Platform evaluation (Google Forms, Moodle LMS, MaxTest, and custom systems) is conducted using both technical and pedagogical criteria. The interdisciplinary integration model proposed here advances the field by bridging formative assessment theory with classroom-level test deployment practices.

Keywords

test software; item design; Bloom's taxonomy; Classical Test Theory; informatics education; ADDIE model; educational technology; assessment integration

1. Introduction

The increasing adoption of digital assessment tools in educational systems worldwide has generated substantial interest in the principled design of test software for classroom and institutional use. Within informatics education, assessment design occupies a particularly important role: the field demands not only factual recall but the evaluation of algorithmic thinking, problem decomposition, and applied computational skills — cognitive competencies that standard psychometric instruments may inadequately capture.

In the Uzbekistan national curriculum, informatics is taught from the middle school level onward, with standardised assessments playing a determinative role in student progression and teacher evaluation. Despite this centrality, the

methodology governing test software design in this context remains undertheorised: practitioners frequently rely on generic survey tools (e.g., Google Forms) with limited psychometric grounding, while research-grade instruments developed abroad (e.g., ETS frameworks, Cambridge Assessment guidelines) are rarely adapted to local didactic conditions.

This paper addresses the gap by presenting a comprehensive, seven-stage methodology for test software development — grounded in ADDIE instructional design, CTT/IRT psychometrics, and Bloom's Revised Taxonomy — and demonstrating its application in an informatics instructional context. Specifically, the study pursues three objectives:

1. To establish a theoretically grounded, step-by-step development model for educational test programs;
2. To propose a principled item classification framework integrating format, cognitive level, difficulty, and assessment purpose;
3. To evaluate available software platforms against pedagogical and technical criteria relevant to informatics instruction.

2. Literature Review

The theoretical foundations of test development draw on two major measurement frameworks. Classical Test Theory (CTT), codified through the work of Brennan (2001), Cronbach (1951), and Millman (1989), defines test reliability and item quality through observed score variance decomposition and item difficulty (p-value) and discrimination (D) indices. Item Response Theory (IRT), developed through contributions by Hambleton, Swaminathan, and Rogers (1991) and extended by van der Linden and Glas (2000), models item performance as a function of latent ability, enabling adaptive testing and more precise difficulty calibration through parameters a (discrimination), b (difficulty), and c (guessing probability).

Bloom's Taxonomy (Bloom, 1956), revised by Anderson and Krathwohl (2001), provides the cognitive classification system most widely used to align instruction, curriculum, and assessment. The six levels — Remember, Understand, Apply, Analyse, Evaluate, Create — form a progressively complex hierarchy that has been shown to structure assessment blueprints across diverse subject domains (Forehand, 2010). Research consistently indicates that a preponderance of test items in formal education target only the lower two levels (remember, understand), while higher-order levels remain underrepresented (Zheng et al., 2008).

The ADDIE model (Analysis, Design, Development, Implementation, Evaluation), originally proposed for instructional systems design, has been adapted for test development contexts by numerous scholars (Dick & Carey, 2005; Molenda, 2003). Its iterative character aligns well with the cyclical nature of psychometric

validation, where pilot testing informs item revision, which in turn feeds back into design decisions. In the domain of informatics education, however, comprehensive ADDIE-grounded test development frameworks remain largely absent from the literature, with most contributions focusing either on curriculum design or platform usability rather than on integrated development methodology.

3. Methodology

3.1 Research Design

This study employs a design-based research (DBR) approach, integrating theoretical framework construction with iterative applied development. The research proceeded in two interlocking phases: (1) a systematic review and synthesis of psychometric, didactic, and instructional design literature to derive a normative development model; and (2) the application of that model to produce and pilot a validated item bank for the Uzbekistan grade 8 informatics curriculum.

3.2 The Seven-Stage Development Model

Drawing on ADDIE, CTT, and IRT, the following seven-stage model is proposed for the systematic development of educational test software:

St	Name	Core Activity	Output
1	Needs ysis	Audience characterisation, ENA, HPT model; .RT objective setting	Technical irements document
2	Design	Test blueprint, Bloom alignment, difficulty ibution	Specification and blueprint
3	Item elopment	Item writing, expert review, distractor analysis	Reviewed item pool
4	Platform tion	Evaluation against technical/pedagogical criteria	Selected software orm
5	Impleme on	System configuration, item upload, interface ng	Deployed test system
6	Pilot ng	Administration to sample (n ≥ 30–50), error ction	Corrected, field-d items
7	Analysis vision	CTT/IRT statistics: p-value, D-index; item bank ate	Validated, revised bank

Table 1. Seven-stage ADDIE-aligned test development model.

The model is explicitly iterative: results from Stage 7 may require return to Stage 3 (item revision), Stage 2 (blueprint adjustment), or even Stage 1 (redefinition of objectives). This recursive architecture is consistent with contemporary

instructional design practice and with the cyclical validation logic of IRT-based item banking.

3.3 Item Classification Framework

Items were classified according to four independent axes: (a) response format (closed-ended: MCQ, true/false, matching, ordering; open-ended: short answer, essay); (b) cognitive level per Bloom's Revised Taxonomy (L1–L6); (c) difficulty tier (easy: $p > 0.70$; moderate: $0.30 \leq p \leq 0.70$; difficult: $p < 0.30$); and (d) assessment purpose (diagnostic, formative, summative, standardised). This multi-axial classification enables systematic construction of balanced test blueprints and facilitates stratified sampling from the item bank for each deployment context.

3.4 Platform Evaluation Criteria

Four platforms were evaluated: Google Forms (Google LLC), Moodle LMS (open-source, Dougiamas, 2002), MaxTest (Uzbekistan national system), and a custom web-based architecture. Evaluation criteria were grouped into two dimensions: technical (item type support, offline capacity, randomisation, proctoring, security architecture) and pedagogical (alignment with national curriculum, feedback quality, analytics depth, accessibility).

4. Results

4.1 Item Quality Standards

Based on CTT analysis principles, a target difficulty distribution of 20% easy ($p > 0.70$), 60% moderate (0.40–0.70), and 20% difficult ($p < 0.40$) is recommended for summative tests, consistent with established psychometric guidance. Discrimination indices of $D \geq 0.30$ are established as the acceptance threshold, with items falling below this threshold flagged for expert revision. Three methodological rules govern item writing quality: (1) language precision — items must be syntactically unambiguous, avoiding double-barrelled constructs and negated stems where possible; (2) distractor plausibility — distractors must represent systematically plausible misconceptions rather than arbitrary foils; (3) structural parallelism — all response options must follow identical grammatical form to avoid construct-irrelevant variance.

4.2 Platform Comparison

Platform	Item Types	Offline	Proctoring	Analytics	Best Fit
Google Forms	10+	No	Limited	Via Sheets	Formative, homework
Moodle LMS	15+ (SCORM, H5P)	No	Browser lock, IP	Advanced reports	HE, distance learning
MaxTest	MCQ,	Yes	Basic	Standard	Schools, DTM

	matching, ordering				prep
Custom system	Unlimited	Configurable	Full (camera, AI)	Custom dashboards	Large-scale, high-stakes

Table 2. Comparative evaluation of test software platforms.

4.3 Integration Framework for Informatics Lessons

The integration of test software into informatics instruction is most effective when conceived not as an administrative add-on but as a component of interdisciplinary, technology-mediated learning design. The interdisciplinary integration model proposed here operates across three functional layers. At the content layer, test items are aligned to curriculum learning outcomes and cross-referenced with adjacent mathematical and scientific concepts, reflecting the longstanding recognition (Ministry of Public Education, Uzbekistan, 2022) that informatics competencies develop in dialogue with mathematical reasoning. At the methodological layer, test delivery is embedded within a formative assessment cycle: diagnostic pre-tests establish baseline competency profiles, formative micro-assessments at the close of each instructional unit provide actionable feedback, and summative end-of-module tests yield data for learning analytics. At the technological layer, platform selection is governed by the technical infrastructure of the deploying institution, with the evaluation framework in Section 3.4 providing decision-support guidance.

This three-layer integration model advances beyond tool-centric approaches by treating assessment not as a discrete event but as a continuous thread woven through instruction. Real-time feedback loops enabled by digital platforms reduce the latency between performance and correction that characterises paper-based assessment, supporting more adaptive instructional responses from teachers.

5. Discussion

The seven-stage model presented in this paper offers several contributions to practice. First, by anchoring the development process in established psychometric theory (CTT difficulty and discrimination indices; IRT parameter estimation), it provides practitioners with quantitative criteria for item quality control rather than relying on expert intuition alone. Second, the explicit linkage of Bloom's Revised Taxonomy to item classification ensures that cognitive coverage is systematically planned rather than incidentally achieved — a particular concern in informatics education, where higher-order skills (analysis, evaluation, creation) are pedagogically central but assessively underrepresented.

Third, the platform evaluation framework operationalises the selection decision in terms of measurable technical and pedagogical criteria, reducing the

influence of marketing claims or institutional inertia. The finding that no single platform dominates across all criteria reinforces the importance of context-specific decision-making: Google Forms remains appropriate for low-stakes formative tasks, while Moodle and custom systems offer the security and analytics infrastructure required for summative and high-stakes applications.

The integration framework extends these findings into the classroom context, positioning test software as an instrument of curriculum-aligned, technology-mediated learning rather than merely an administrative assessment tool. This reframing is consistent with broader movements in educational technology toward the blending of learning and assessment, sometimes described under the concept of assessment as learning (Earl, 2003).

Limitations of this study include the contextual specificity of the deployment setting (Uzbekistan grade 8 informatics) and the absence of a large-scale empirical pilot producing full IRT parameter estimates. Future research should conduct large-sample calibration studies ($n \geq 200$) using Rasch or 2PL models, examine the platform integration model's effectiveness through quasi-experimental designs, and explore the applicability of the framework to other STEM disciplines within the national curriculum.

6. Conclusion

This paper has established a seven-stage, psychometrically grounded methodology for the development and classroom integration of educational test software in informatics instruction. The proposed framework — anchored in CTT, IRT, and Bloom's Revised Taxonomy, structured through the ADDIE model, and operationalised through a multi-axial item classification system — addresses a documented gap between the theoretical sophistication of assessment science and the practical realities of test software deployment in national education systems.

The comparative platform evaluation and three-layer integration model provide decision-support tools applicable across diverse institutional contexts. The study contributes to the growing body of literature on evidence-based assessment design in computer science education and offers a replicable methodological template for curriculum developers, educational technologists, and classroom practitioners seeking to harness digital assessment tools in principled, pedagogically coherent ways.

REFERENCES:

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

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay.

Brennan, R. L. (2001). Generalizability theory. Springer.

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

Dick, W., Carey, L., & Carey, J. O. (2005). The systematic design of instruction (6th ed.). Allyn & Bacon.

Earl, L. M. (2003). Assessment as learning: Using classroom assessment to maximize student learning. Corwin Press.

Forehand, M. (2010). Bloom's taxonomy. In M. Orey (Ed.), Emerging perspectives on learning, teaching, and technology. University of Georgia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. SAGE Publications.

Millman, J. (1989). If at first you don't succeed: Chess, basketball, and the standardised test. Phi Delta Kappan, 71(3), 223–228.

Molenda, M. (2003). In search of the elusive ADDIE model. Performance Improvement, 42(5), 34–36. https://doi.org/10.1002/pfi.4930420508

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). Computerized adaptive testing: Theory and practice. Kluwer Academic Publishers.

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's taxonomy debunks the 'MCAT myth'. Science, 319(5862), 414–415. https://doi.org/10.1126/science.1147852