IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.1511165

From Student Performance Analytics to Personalized Assessment: A Survey of Learning Analytics and Automated Test Generation

Abhishek Bobade[1], Tejas Chaudhari[2], Yash Bansode[3], Dr. Archana Ghotkar[4]

Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India[1,2,3]

Associate Professor, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology Pune, India[4]

Abstract: Online mock-test platforms accumulate rich response histories, yet a single aggregate score says little about what a learner should study next. This survey examines how assessment logs can be turned into chapter-level evidence, and how that evidence can steer the construction of practice papers that must still follow a fixed examination blueprint. We review descriptive learning analytics, uncertainty-aware proportion estimates, item response theory, cognitive diagnosis, knowledge tracing, automated test assembly, shadow testing, computerized adaptive testing, and the treatment of infeasible assembly models. The families are compared on their inputs, outputs, assumptions, data demands, interpretability, and capacity to enforce structural constraints. We formalise one way of coupling the two stages, discuss evaluation practice, and identify open research gaps. The article is a scoped narrative survey; it is neither a systematic review nor a report of empirical learning gains.

Index Terms: Learning analytics, educational data mining, chapter-level analytics, automated test assembly, exam blueprint, computerized adaptive testing

1. INTRODUCTION

Large-scale entrance examinations, for example JEE Main and MHT-CET in India, are commonly approached through repeated timed mock papers. A candidate who finishes such a paper typically receives a mark and perhaps a rank, but little insight into where the marks were lost, whether the losses recur, or what the following paper should emphasise. An acceptable total may hide a few chapters that are missed in every sitting, and a weak total may stem from poor time management rather than conceptual gaps.

Digital delivery changes what can be known about a learner. Each attempt can be stored together with the identity of every question, the option chosen, the outcome, the time spent, and the subject, chapter and difficulty labels attached to the question in the item bank. Two requirements follow from this. First, the platform should report, with honest evidence, how each chapter is progressing. Second, it should let that knowledge influence the next paper without departing from the structure of the real examination.

The research that speaks to these requirements is spread over largely separate communities. Learning analytics has examined how performance data are presented to learners [1]–[4], whereas psychometrics and operations-oriented work have examined how papers are assembled from item banks under strict specifications [6]–[8], [15]. This survey therefore follows the entire route from raw responses to the next paper and asks what each stage silently assumes about the one before it.

A. Related Surveys and Positioning

Several existing reviews cover parts of this territory. Early and updated surveys of educational data mining and learning analytics map the method families and the shift from purely technical mining toward stakeholder-facing analytics [9], [10]. Reviews of learner-facing dashboards and open learner models document what is displayed and how learners react [1]–[3], and a broad review of learning analytics in higher education reports that most studies remain descriptive and few demonstrate improved outcomes [11]. The test-assembly and adaptive-testing literature is reviewed in monographs and in comparison studies of formulations and solvers [6], [26], [27]. The present article takes a pipeline view, tracing how a chapter-level estimate produced by analytics can become an objective or a constraint in assembly.

B. Contributions

The main contributions are as follows.

- We present a reference pipeline for analytics-driven practice generation (Section 3) and use it as a common frame for comparing otherwise unrelated methods.

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.1511165

- We organise chapter-level performance analytics (Section 4) around accuracy, recency, interval and shrinkage estimates, and evidence levels that separate "weak" from "not yet known".
- We organise the methods that connect analytics to test generation (Section 5), formulate chapter-priority selection under blueprint constraints, distinguish local validation failures from global infeasibility, and describe a repair-and-fallback policy.
- We compare method families on shared criteria, consolidate evaluation practice, and identify research gaps (Section 6).

C. Scope

The survey covers work that analyses assessment response data and uses it to decide what a learner practises next. Hardware, remote proctoring, automatic question authoring and seat-allocation modelling lie outside its scope. Since the target setting is dominated by multiple-choice and numerical-answer items, the quantitative material concentrates on dichotomously scored responses, and extensions to partial credit are mentioned where relevant. A study is included when it contributes a method, a design principle or an evaluation finding concerning learner modelling, dashboard-based feedback, item selection, test assembly or assessment constraints. Fig. 1 shows how the paper is organised.

D. Review Approach and Research Questions

This is a focused narrative survey and not a protocol-driven systematic review. The cited corpus combines foundational work on educational data mining, learner modelling and test assembly with later studies on learning analytics dashboards, adaptive testing and optimisation-based assessment. Studies are grouped by the role they play in the response-to-practice pipeline and compared qualitatively, with no pooled effect estimate. Database hit counts, duplicate-removal figures and a PRISMA flow are not reported, because no reproducible multi-database screening protocol was carried out. The review should therefore be read as a scoped synthesis of representative methods and not as an exhaustive census of the field.
Three research questions guide the discussion. **RQ1:** How can chapter-level performance be summarised without overstating the evidence when response counts are small? **RQ2:** How can learner-specific priorities enter paper generation while the examination constraints are preserved? **RQ3:** Which design and evaluation requirements are needed before the resulting practice can be called valid, explainable and educationally useful?

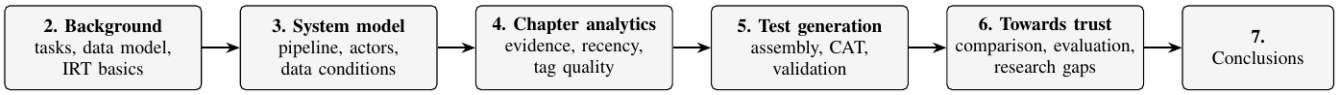


Fig. 1. Outline of this survey. Sections 4 and 5 follow the data from per-chapter evidence to the choice of the next paper.

2. About Assessment Analytics

A. Terminology and Tasks

Educational data mining (EDM) and learning analytics (LA) overlap considerably. A common distinction regards the former as method-oriented and the latter as stakeholder-oriented [10], and early reviews already charted the principal method families [9]. Practice platforms draw on both, because an estimate has little value to a candidate unless an interface explains it.
A *learning analytics dashboard* presents such estimates visually. Dashboards intended for learners are closely related to open learner models, which expose what a system believes about the learner [1], [2]. A *blueprint* is the specification of an examination paper: the number of questions, their allocation to subjects, the expected syllabus coverage and the intended difficulty mix. Blueprints differ between examinations, so a generator aimed at JEE Main or MHT-CET has to treat them as parameters. An *evidence level* is a label that states how much data support a chapter-level claim.
Three tasks recur throughout the literature. *Summarisation* reduces interactions to counts and rates per chapter. *Estimation* infers a learner's latent proficiency on a concept or overall. *Selection* decides which questions to present, given the estimates and any structural requirements.

B. Data Model and Estimation Basics

For learner u, a platform stores a sequence of interactions (q_t, c_t, r_t, τ_t): the question, its chapter label, the outcome (correct, incorrect, skipped, or a graded mark) and a timestamp. Analysis can be performed on a single response, on a chapter within one attempt, on an entire attempt, or on the longitudinal history across attempts.
Item response theory (IRT) models the probability of a correct response through a latent ability and item parameters. In the one-parameter Rasch form,

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.1511165

$$P(r = 1 \mid \theta, b) = \frac{1}{1+e^{-(\theta-b)}}, \qquad (1)$$

with ability θ and item difficulty b; richer forms add discrimination and guessing parameters. IRT has been applied to match learning material to learners [13], [14]. Cognitive diagnosis estimates proficiency per concept through a Q-matrix recording which concepts each item assesses [12], while knowledge tracing follows proficiency over a sequence of practice opportunities [23], [24]. All of these approaches presuppose correct item labelling and, in the case of IRT, calibrated item parameters. When the items of a bank are labelled by hand and rarely administered, neither condition can be taken for granted.

3. System Model

To compare methods on equal terms, the discussion is organised around the reference pipeline in Fig. 2. It is a deliberately generic architecture: any practice platform that stores item metadata and attempt logs can be mapped onto it.

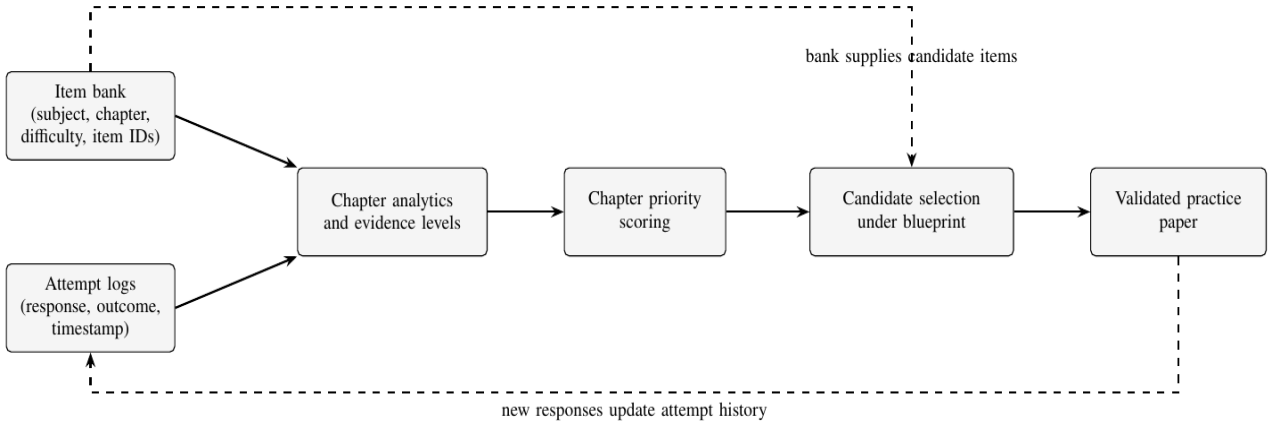


Fig. 2. Reference pipeline. Attempt logs and item metadata jointly support chapter analytics; the item bank also supplies candidate questions to the constrained selection stage.

A. Pipeline and Failure Surface

Attempt logs and item metadata are joined through stable item identifiers. The logs contribute learner responses and timestamps, and the item bank contributes subject, chapter and difficulty attributes. The joined records are aggregated per chapter, converted into evidence-aware priorities, and used to rank candidate questions. Selection must then satisfy the configured blueprint before any paper is released. Errors can enter at every hand-over. A missing or incorrect chapter tag corrupts a summary; a rate computed from two attempts is mistaken for a trend; a priority list requests more questions from a chapter than the bank contains; and a ranking that ignores the blueprint yields a paper unlike the real one. The remaining sections read the literature for the way each hand-over is protected.

B. Actors

Four groups interact with such a system. *Candidates* consume dashboards and papers. *Instructors or curators* maintain the bank and its tags. *Maintainers* configure thresholds and blueprint parameters. *Examination bodies* define the paper pattern that the platform imitates but do not take part in its operation. Each group needs a different kind of explanation: candidates need plain labels accompanied by evidence counts, curators need audit views of tag consistency, and maintainers need to understand why a paper was accepted or rejected.

C. Data Conditions

Four conditions shape every method discussed below. *Sparse evidence:* a candidate may have answered only a handful of questions per chapter. *Label noise:* manually assigned chapter and difficulty tags are imperfect. *Thin banks:* a chapter may lack sufficient questions of a given difficulty. *Self-selection:* learners choose how much to practise and when, which confounds before-and-after comparisons.

D. Design Objectives

Five objectives are used to judge the methods: (i) *honest evidence*, meaning that a statement never appears more certain than its data allow; (ii) *structural validity*, meaning that generated papers respect the blueprint; (iii) *responsiveness*, meaning that the paper genuinely changes when the candidate's profile changes; (iv) *auditability*, meaning that a reviewer

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.1511165

can see why an item was chosen; and (v) *safe failure*, meaning that an impossible request ends in a standard valid paper and not in a broken one.

4. CHAPTER-LEVEL PERFORMANCE ANALYTICS

A. Dashboards and Descriptive Products

The most common output of a practice platform is a table of counts and rates per chapter: attempted, correct, incorrect, skipped, and how accuracy has moved over recent attempts. Systematic reviews describe what learner-facing dashboards show and how learners respond to them [1], [2]; empirical studies have been analysed through a self-regulated learning lens [3]; and a checklist now spans planning, design, implementation and evaluation [4]. A chapter view for examination preparation is one instance of this broader design space. Findings from teacher-facing dashboard studies can inform the design, but their transfer to learner-facing examination tools is an adaptation and should not be read as direct evidence. A sobering result comes from a review of 252 higher-education studies, in which most work was descriptive and only about 9% reported improved learning outcomes [11]. Showing data is therefore not the same as supporting learning.

A critical reading of these sources reveals two weaknesses. First, dashboards are usually evaluated through usability or satisfaction measures, so little is known about whether a chapter label changes what a candidate studies. Second, the statistical status of the displayed numbers is rarely examined; a rate shown without its count carries no information about its reliability. The next subsection addresses the second point directly.

Key observation: Descriptive chapter views are cheap and widely deployed, yet their benefit is seldom demonstrated. They should be designed and evaluated with published dashboard guidance and not built ad hoc.

B. Quantifying Evidence

For chapter c, let n_c^{pres} be the number of scored opportunities presented, k_c the number answered correctly, and s_c the number skipped. For an exam-readiness accuracy measure, we define a skipped item as a non-correct opportunity and use $a_c = k_c/n_c^{\text{pres}}$ when $n_c^{\text{pres}} > 0$. We report the answer rate $(n_c^{\text{pres}} - s_c)/n_c^{\text{pres}}$ separately so that skipping is not hidden inside the accuracy value. If the intended question is accuracy conditional on answering, the denominator must instead be the number answered; the two measures must not be conflated. For compactness, write $n_c = n_c^{\text{pres}}$ in the remaining chapter-level formulas. A chapter with no presented items is shown as "no data", never as 0%. Small denominators are also problematic: one correct answer out of two opportunities yields 50% with substantial uncertainty. Two useful remedies are intervals and shrinkage. Where an examination uses negative marking or partial credit, the platform should additionally report the official score metric rather than treating binary accuracy as a substitute for marks.

The Wilson score interval [20] is recommended over the plain normal interval for proportions with small samples [21]. Its half-width is approximately

$$h(n) \approx \frac{z\sqrt{a(1-a)/n + z^2/(4n^2)}}{1 + z^2/n}, \qquad (2)$$

with $z = 1.96$ for 95% coverage. At $a = 0.5$ it equals about 0.41 for $n = 2$, 0.26 for $n = 10$ and 0.13 for $n = 50$ (Fig. 3). A "weak" label drawn from a handful of questions is therefore statistically fragile.

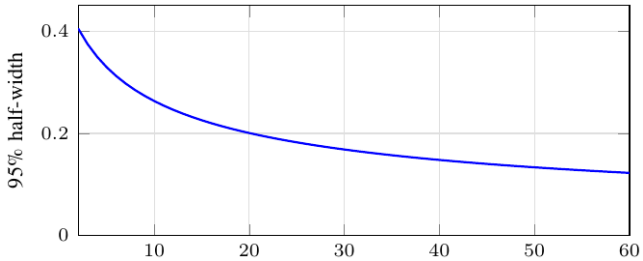


Fig. 3. Half-width of the Wilson 95% interval at an observed accuracy of 50%, as a function of attempted questions.

Shrinkage can stabilise a chapter estimate by combining its observations with a prior mean a_0, for example a cohort or learner-level baseline. The following beta-binomial posterior-mean form is a simple smoothing rule; it is empirical Bayes only when its prior parameters are estimated from a population of learners or items [22]:

$$\tilde{a}_c = \frac{k_c + m a_0}{n_c + m}, \qquad \varepsilon_c = \frac{n_c}{n_c + m}, \qquad (3)$$

where $m > 0$ is the prior strength and $\varepsilon_c \in [0, 1)$ is an evidence weight that grows with the number of presented opportunities. The values of a_0 and m should be estimated or justified on development data and checked for sensitivity;

DOI: 10.17148/IJARCCE.2026.1511165

they are not universal constants. Intervals communicate uncertainty, whereas shrinkage yields a stabilised point estimate for ranking; both should be shown when space permits.

An *evidence level* turns counts into a label. Table I gives an illustrative scheme of the kind found in practice dashboards; the cut-offs are configuration values that should be set from data, not constants of the method.

TABLE I
ILLUSTRATIVE EVIDENCE LEVELS FOR A CHAPTER SUMMARY.

Level	Condition	Display rule
No data	$n_c = 0$	"Not attempted"; no rate
Insufficient	$0 < n_c < n^L_{\min}$	Counts only; no weak/strong label
Low	$n^L_{\min} \le n_c < n^A_{\min}$	Rate with interval; tentative wording
Adequate	$n_c \ge n^A_{\min}$	Rate, interval and label

Key observation: Reporting a rate without its count invites overconfident labels. Intervals, shrinkage and explicit evidence levels are inexpensive safeguards, and they make "weak" distinguishable from "unknown".

C. Recency and Trend

Recent behaviour is more informative than old behaviour. Two common devices are a fixed window over the last w attempts and exponentially decayed weights. With decay factor $\lambda \in (0, 1)$ and per-attempt counts (k_j, n_j) for $j = 1, \ldots, T$, where n_j counts presented opportunities and skipped items are non-correct as defined above,

$$a_c^{\text{rec}} = \frac{\sum_{j=1}^{T} \lambda^{T-j} k_j}{\sum_{j=1}^{T} \lambda^{T-j} n_j}. \qquad (4)$$

Recent figures should carry their own sample size, since a window of three attempts may contain very few questions from a given chapter. The choice of w or λ is seldom justified in published dashboards; it can be tuned by checking how well recent accuracy predicts accuracy on later attempts. The same logic connects to the memory literature: retrieval practice and spacing both affect retention [28], [29], so a decay rate that is too aggressive may discard evidence that is still relevant, while one that is too slow may mask genuine improvement.

D. Tag Quality

Chapter analytics are only as reliable as the tags beneath them. A question without a chapter label cannot contribute to a summary, and a mislabelled one contaminates two chapters at once. The Q-matrix of cognitive diagnosis makes this dependence explicit [12]. For examinations with published syllabi, curated tags are realistic, so quality control at import time (every question must carry subject, chapter and difficulty, with no duplicates) is the practical priority. Empirical difficulty from IRT can later refine expert labels once responses accumulate [13], although a newly added item has no history to calibrate from.

Key observation: Tag validity is a precondition and not a detail. When chapter lists are fixed, import-time checks and later consistency audits matter more than model sophistication.

5. FROM ANALYTICS TO TEST GENERATION

A. Ability-Matched Selection

The simplest use of an ability estimate is to present items whose difficulty lies close to the learner's current level, as in IRT-based personalised e-learning [13] and in the item-based adaptive environments reviewed by Wauters *et al.* [14]. This presumes calibrated items and treats questions independently, so nothing guarantees that the resulting set resembles a real paper. The approach also optimises for measurement or engagement, not for rehearsal of a particular exam format.

B. Assembly as Constrained Selection

A complementary body of research treats test construction as constrained optimisation. Theunissen formulated test design as a binary program [25], and van der Linden later gave a systematic treatment of linear models for optimal test design [26]. Building on this foundation, Hwang *et al.* studied test-sheet composition from large item banks, accounting for mean difficulty, discrimination, testing time, item count and concept weights. They related the selection problem to the NP-hard knapsack problem and proposed algorithms that produce near-optimal papers within reasonable computation time [15]. Later work showed that solution quality can vary with the mathematical formulation of the constraints and with the solver [6]. Related research considered the construction of adaptive item pools as collections of linear tests [5]. Genetic

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2026.1511165

algorithms offer another approximate route and allow flexible objective functions, but their stochastic search brings additional tuning and reproducibility concerns [16].

C. Shadow Tests and Adaptive Testing

In shadow-test methods a complete test satisfying all constraints is assembled before each item is chosen, so every administered test stays within specification, including in variable-length settings [7]. Computerized adaptive testing (CAT) has a long theoretical tradition [27], and more recent learned selection policies include BOBCAT, which learns the selection rule from data through bilevel optimisation [17], neural CAT designed for online education [18], and a formative-assessment system that couples CAT with a learning memory cycle [19]. These approaches choose each item from the running ability estimate in order to measure or support the learner; they generally do not impose subject counts, syllabus coverage and a difficulty mix on a complete fixed-form paper.

Key observation: Assembly methods are built to guarantee structure, whereas adaptive methods are built to track the learner. The two properties are normally obtained from different procedures.

D. Priority-Driven Selection under a Blueprint

Personalisation and blueprint fidelity pull in opposite directions: favouring weak chapters moves the paper away from the standard distribution, while enforcing the standard distribution leaves little room to personalise. A clean way to state the trade-off is to make chapter priority the objective and the blueprint the constraint set. Let $x_i \in \{0, 1\}$ indicate selection of question i and $\pi_{c(i)}$ the priority of its chapter:

$$\begin{aligned} &\max \textstyle\sum_i \pi_{c(i)} x_i \quad \text{s.t.} \quad \textstyle\sum_i x_i = N, \\ &\textstyle\sum_{i \in S} x_i = N_s \; \forall s, \quad L_d \le \textstyle\sum_{i \in D} x_i \le U_d \; \forall d, \\ &\textstyle\sum_{i \in C} x_i \ge m_j \; \forall j, \quad \textstyle\sum_{i \in P} x_i \le \rho N, \end{aligned} \tag{5}$$

where N is the total number of questions; S_s is the set of items in subject s with required count N_s; D_d is the set in difficulty class d with bounds $[L_d, U_d]$; C_j is the set tagged to chapter j with minimum coverage m_j; and P is the set of items from chapters currently prioritised for the learner. The binary variables ensure that an item is selected at most once. The official examination specification supplies hard requirements such as total count, subject allocation and any published difficulty or syllabus constraints. Personalisation parameters such as m_j and the priority-share cap ρ are policy choices of the platform and must not be presented as official examination rules unless the examining body specifies them. These values should be configured separately for each examination and checked against the available bank. A natural priority uses both weakness and evidence:

$$\pi_c = \varepsilon_c (1 - \tilde{a}_c), \tag{6}$$

which rises when a chapter is weak *and* well supported, and stays low when a poor rate rests on very few questions.

E. Infeasibility and Constraint Relaxation

A candidate paper can fail validation even when a feasible paper exists; this is a local failure of the selected set, not proof that the full assembly model is infeasible. Global infeasibility means that no item set in the available bank satisfies all currently active constraints. The distinction matters because a validator that checks one ranked paper cannot establish global infeasibility. A solver, a complete search procedure, or a documented feasibility analysis is required for that conclusion [8].

We therefore separate constraints into two classes. Hard constraints reproduce the official paper requirements and must not be relaxed silently. Soft personalisation constraints include chapter-priority caps and optional chapter minima introduced to shape the learner's practice. If the personalised request is infeasible, the system first removes or relaxes soft personalisation constraints in a documented order and re-runs feasibility checking. If a paper satisfying the hard blueprint still cannot be assembled from the bank, the system must not release an invalid paper: it should report a bank or configuration failure to the maintainer. A standard fallback paper is safe only when it has been independently assembled and validated against the same hard blueprint.

F. Rank-then-Validate Selection

Three design families are relevant to priority-driven generation: exact optimisation, heuristic or evolutionary search, and priority ranking followed by constrained selection and validation. The third family is discussed here as a design pattern and not as an empirically benchmarked algorithm. A naive top-N ranking is insufficient, because it can violate subject, difficulty or chapter constraints. The practical variant must pick high-priority items while checking remaining quota and bank capacity at each step, and then attempt bounded swaps or reselection when a check fails.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.1511165

The procedure is: (1) compute chapter priorities from the learner's evidence-aware profile; (2) rank eligible items using chapter priority, with item quality and exposure limits as secondary criteria; (3) construct a candidate paper while reserving capacity for unmet hard quotas; (4) validate total count, subject allocation, difficulty bounds, chapter requirements and item integrity; (5) attempt deterministic swaps or re-selection for any violation; and (6) release the personalised paper only after all hard checks pass. If the personalised constraints remain infeasible, relax only the documented soft personalisation rules and retry. If the hard blueprint cannot be satisfied, do not release a paper; flag the bank or configuration for correction. A standard fallback may be released only if it is already known to satisfy the hard blueprint. Fig. 4 summarises this control flow.

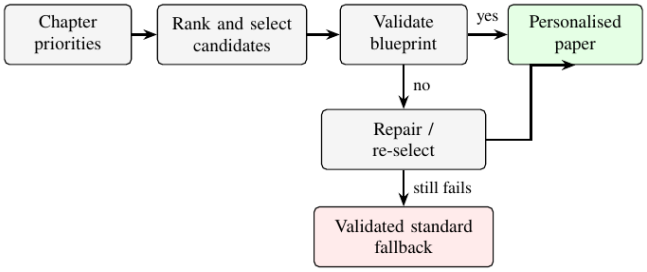


Fig. 4. Rank, constrained selection, validation and repair. An unresolved personalised request falls back only to a standard paper independently validated against the hard blueprint.

TABLE II
VALIDATION AND RECOVERY RULES FOR A CANDIDATE PAPER.

Check	Rule	Action on failure
Question count	Total selected equals N	Re-select items
Subject mix	Each subject count equals N_s	Swap within quota
Difficulty	Each class lies in $[L_d, U_d]$	Swap by class
Coverage	Required chapter minima are met	Fill available gaps
Priority cap	Priority share is at most ρ	Relax soft cap if needed
Integrity	No duplicates; required tags exist	Reject or repair item
Local validity	Current candidate violates a hard rule	Repair and recheck
Global feasibility	No feasible set exists in the bank	Feasibility analysis; alert maintainer
Fallback	Prevalidated standard paper exists	Release only after recheck

Key observation: Keeping official blueprint requirements hard while treating personalisation rules as configurable soft constraints makes the design portable across examinations. A fallback is safe only when it has independently passed the hard blueprint checks.

6. TOWARDS TRUSTWORTHY ADAPTIVE PRACTICE

A. Comparative Analysis

Table III compares method families on input, output, key assumptions, interpretability, data demand and constraint handling. The qualitative labels are an analytical synthesis and not measured scores: "high" interpretability means that a selection or estimate can be traced to explicit counts, parameters or constraints; "medium" means that some model behaviour requires explanation; and "low" means that the decision depends substantially on learned representations or policies. Data demand is relative to the calibration and training needs of each method, not a universal sample-size

DOI: 10.17148/IJARCCE.2026.1511165

threshold. No family dominates. Transparent summaries depend on reliable tags and sufficient observations; learned adaptive methods generally require calibration or training data; and assembly methods can enforce structural constraints but do not automatically use learner-level chapter evidence.

Reading the table column by column clarifies several trade-offs. Along the interpretability axis, descriptive summaries and integer programs sit at the transparent end, whereas neural CAT and deep knowledge tracing sit at the opaque end; this matters for the auditability objective of Section 3. Along the constraint axis, only integer programming, shadow tests and rank-then-validate designs provide explicit guarantees. Along the data axis, the methods that offer the richest learner models (cognitive diagnosis, knowledge tracing, neural CAT) are also the most data hungry, which is a poor match for learners who attempt a few dozen papers at most.

TABLE III
COMPARISON OF METHOD FAMILIES ON COMMON CRITERIA.

Family	Input	Output	Key assumption	Interpret.	Data need	Constraints
Descriptive dashboards	Tagged responses	Counts, rates, trends	Correct tags; enough attempts	High	Low	None
Interval / shrinkage	Tagged responses, prior	Adjusted rate, evidence level	Sensible prior and thresholds	High	Low	None
IRT, diagnosis, tracing	Responses, item parameters, Q-matrix	Ability or concept estimates	Model fit; item/tag quality	Medium–high	Medium–high	None by default
Ability-matched selection	Ability, calibrated items	Next items	Calibration and exposure control	High–medium	Calibration data	Not guaranteed
Integer programming	Bank with attributes	Assembled paper	Calibrated attributes; feasibility	High	Item bank	Explicit
Shadow-test CAT	Bank, ability, constraints	Constrained shadow form	Feasible calibrated pool	Medium	Calibrated bank	Explicit
Learned / neural CAT	Response history, bank	Next item	Enough data	Low–medium	High	Usually none
Genetic search	Bank	Assembled paper	Fitness encodes needs	Medium	Item bank	Penalty or repair
Rank then validate (hybrid)	Priorities, bank, blueprint	Paper or safe fallback	Complete checks and repair policy	High	Low–medium	Explicit checks; feasibility separate

TABLE IV
REPRESENTATIVE STUDIES GROUPED BY TASK. ENTRIES SUMMARISE WHAT THE PAPERS REPORT; THEY ARE NOT A HEAD-TO-HEAD BENCHMARK.

Ref.	Year	Task	Technique	Reported focus or finding
[25]	1985	Test construction	Binary programming	Test design as optimisation
[23]	1994	Student modelling	Knowledge tracing	Tracks mastery across practice
[16]	2005	Test-sheet generation	Genetic algorithm	Embedded in a testing system
[13]	2005	Personalised e-learning	Rasch IRT	More effective learning than a non-personalised path
[8]	2005	Infeasible assembly	Method comparison	Compares ways to handle infeasibility
[5]	2006	Adaptive pool assembly	Linear tests	Pool built as a set of linear tests
[15]	2006	Test-sheet composition	Mixed integer programming	Near-optimal sheets in reasonable time
[14]	2010	Adaptive learning	IRT-driven selection	Possibilities and challenges
[24]	2015	Knowledge tracing	Recurrent networks	Learned model of student knowledge
[11]	2018	Review of LA	—	Mostly descriptive; about 9% show improved outcomes
[7]	2018	Variable-length CAT	Shadow tests	Constructing valid shadow tests
[2]	2018	Review of dashboards	—	Open learner models and dashboards
[3]	2020	Review of dashboards	—	Self-regulated learning perspective
[6]	2020	Automated assembly	MIP	Modelling and solver affect outcomes

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.1511165

Ref.	Year	Task	Technique	Reported focus or finding
[12]	2020	Cognitive diagnosis	Neural, monotonicity	Accurate, interpretable proficiency
[17]	2021	Adaptive testing	Bilevel optimisation	Learns selection from data
[1]	2021	Review of dashboards	—	Learner-facing dashboards
[18]	2022	Adaptive testing	Neural CAT	Fully adaptive selection for online education
[19]	2022	Formative assessment	CAT with memory cycle	Personalised learning support
[4]	2023	Teacher-facing dashboard evaluation	Checklist	Planning-to-evaluation guidance; transfer to learner dashboards requires care

B. Evaluation Practice

Analytics as interface and as statistic. As interfaces, dashboards can be assessed against the checklist of Kaliisa *et al.* [4] and, where learners participate, against the self-regulated learning view [3]. As statistics, an evidence threshold can be audited for stability: a chapter flagged weak on three questions that no longer looks weak after ten suggests that the minimum count is too low.

Generation validity. Useful measures include the proportion of candidate papers that pass every hard constraint, the number and type of repairs, the fallback frequency, item-bank coverage, the duplicate rate and the generation time. If an exact solver is used, the feasibility status, objective value and solver gap or termination status should be reported; a heuristic should not be described as proving infeasibility. Such measures separate structural correctness from personalisation quality [6], [8].

Personalisation quality. Relevant indicators are the share of selected questions drawn from priority chapters, the distribution of selected difficulty, the exposure of individual items, and the difference between a learner-specific paper and the standard blueprint paper. These should be reported alongside hard-constraint pass rates, so that a higher priority score is not mistaken for a better test.

Learning outcomes. The decisive test is performance on a later, comparable paper, preferably complemented by delayed retention or transfer measures. Before-and-after differences alone cannot separate personalisation from ordinary practice effects, since retrieval practice improves retention [28] and spacing influences it [29]. A prospective study should compare analytics-guided practice with a fixed-blueprint control, randomise learners where feasible (or use matched groups), use comparable outcome tests, report baseline differences and attrition, and predefine its primary outcomes. Until such studies are available, the educational benefit of analytics-guided blueprint-constrained practice remains a hypothesis and not an established result.

Key observation: Existing evidence mostly shows that requirements are met or that learners were satisfied in bounded studies. Controlled evidence that analytics-guided papers improve later performance is scarce.

C. Privacy and Responsible Use

Response histories are personal data and may reveal educational strengths, weaknesses and study habits. A platform should define a lawful purpose and retention period, collect only necessary fields, protect data in transit and at rest, restrict access by role, and provide a process for correction or deletion consistent with applicable institutional policy and law. Reports should show counts and uncertainty so that a candidate is not discouraged by an unsupported “weak” label. Priorities should be explainable: a candidate or instructor should be able to see which chapter signals affected selection and which blueprint constraints limited the result. Evaluation datasets should be de-identified where practical, with access to re-identification keys tightly controlled.

D. Research Gaps

1) Lack of Integration Between Chapter-Level Analytics and Blueprint-Constrained Test Generation:

The literature reviewed above falls into two groups whose outputs do not fit together. Analytics work produces per-chapter descriptions of a learner, such as accuracy, recency-weighted rates, intervals and evidence labels, and in most systems these end at the dashboard. Assembly work produces papers that satisfy counts, difficulty bounds and coverage rules, but the quantity being optimised is normally a property of the test, for example target information or mean difficulty, and not of the person who will sit it. Each group was designed to answer a different question, so a chapter rate is not directly usable as an item attribute or as the scalar objective that an assembly model expects.

The consequences show up in practice. If chapter estimates are converted to priorities and used to rank items, the ranking knows nothing about quotas, and the paper that results may break the examination pattern. Repairing it can then remove the very emphasis that was intended. If instead the blueprint is applied first and personalisation is added afterwards, little freedom may remain to reflect the learner’s needs. How much freedom exists depends on the bank, since a chapter with



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.1511165

few items of a given difficulty can make a requested emphasis impossible to satisfy. We are not aware of studies that quantify this tension, for instance by measuring how far a personalised paper moves from the standard distribution, how often repair is needed, or how both change with bank size.

Adaptive testing does not settle the matter. Its selection rules aim at measurement precision or learning support, not at rehearsing a fixed paper pattern [17]–[19]. Shadow-test methods come closest, because every administered form stays feasible while the selection adapts [7], but they presuppose calibrated pools and reliable ability estimates, which a practice setting with hand-assigned tags and short response histories seldom has. What is still missing is an accepted formulation in which uncertain chapter evidence supplies the objective, the examination blueprint supplies the constraints, infeasible requests follow a defined relaxation order [8], and results are judged on both structural validity and personalisation. Formulation (5) is one candidate, but it has not been validated against alternatives such as exact solvers [6] or heuristic search [16].

2) Limited Integration of Score Prediction with Personalized Practice Recommendations:

Practice platforms usually attempt two jobs separately: estimating how a candidate is likely to score in the real examination, and advising what to practise next. The first is typically a number derived from recent totals, and the second a list of weak chapters. The two seldom share a model. Learner-modelling methods do predict the chance of a correct response [12], [13], [23], [24], but they are generally evaluated at the level of single responses and not as marks on a full paper with a given mark distribution.

A simple formulation shows what a joint treatment would involve. Suppose chapter c carries a share W_c of the total marks, with $\Sigma_c\ W_c = 1$ after normalisation. Using the stabilised chapter estimates of Section 4, an expected score and its approximate spread are

$$S = \sum_c W_c \tilde{a}_c\,, \qquad \mathrm{Var}(S) \approx \sum_c W_c^2\, \frac{\tilde{a}_c(1-\tilde{a}_c)}{n_c+m}, \qquad (7)$$

where the variance expression assumes independent chapter estimates and ignores negative marking. If practice raises the accuracy of chapter c by Δ_c, the expected gain is $W_c \Delta_c$. A recommendation consistent with the prediction should therefore weigh how many marks a chapter carries and how much room it has to improve, instead of looking at accuracy alone. A chapter at 40% accuracy that contributes few marks may deserve less time than one at 60% that contributes many. Weak-chapter lists ignore W_c, while score predictors ignore how practice changes $\tilde{a}_c$.

Other factors widen the gap. Score depends on time allocation, question-selection strategy and negative marking, none of which appear in a chapter accuracy. The uncertainty in S should be shown with the prediction, since a bare figure can discourage a candidate or encourage overconfidence, and there is little evidence on how learners react to either. Finally, predictions are rarely checked against actual examination results, because linking mock-test records to official outcomes is difficult and raises privacy concerns. Until such validation exists, any claim that a predicted score helps to choose practice remains unverified.

7. Conclusions

This survey connected chapter-level response analytics to blueprint-constrained test generation. Evidence-aware summaries address sparse responses, and assembly methods preserve the structure of an examination. The central design challenge is to personalise selection without relaxing the official constraints. A pipeline with explicit validation, separate handling of local and global infeasibility, and a safe fallback is one workable response, although its learning benefit has not been tested. Two research gaps remain: the weak coupling between chapter analytics and constrained paper generation, and the separation of score prediction from practice recommendation.

References

[1] N. Valle, P. Antonenko, K. Dawson, and A. C. Huggins-Manley, "Staying on target: A systematic literature review on learner-facing learning analytics dashboards," *British J. Educational Technology*, vol. 52, no. 4, pp. 1724–1748, 2021.

[2] R. Bodily *et al.*, "Open learner models and learning analytics dashboards: A systematic review," in *Proc. 8th Int. Conf. Learning Analytics and Knowledge (LAK)*, 2018, pp. 41–50, doi: 10.1145/3170358.3170409.

[3] W. Matcha, N. A. Uzir, D. Gašević, and A. Pardo, "A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective," *IEEE Trans. Learning Technologies*, vol. 13, no. 2, pp. 226–245, 2020.

[4] R. Kaliisa, I. Jivet, and P. Prinsloo, "A checklist to guide the planning, designing, implementation, and evaluation of learning analytics dashboards," *Int. J. Educational Technology in Higher Education*, vol. 20, Art. no. 28, 2023, doi: 10.1186/s41239-023-00394-6.

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2026.1511165

[5] W. J. van der Linden, A. Ariel, and B. P. Veldkamp, "Assembling a computerized adaptive testing item pool as a set of linear tests," *J. Educational and Behavioral Statistics*, vol. 31, no. 1, pp. 91–100, 2006, doi: 10.3102/10769986031001081.

[6] X. Luo, "Automated test assembly with mixed-integer programming: The effects of modeling approaches and solvers," *J. Educational Measurement*, vol. 57, no. 4, pp. 547–565, 2020, doi: 10.1111/jedm.12262.

[7] Q. Diao and H. Ren, "Constructing shadow tests in variable-length adaptive testing," *Applied Psychological Measurement*, vol. 42, no. 7, pp. 538–552, 2018.

[8] H. A. Huitzing, B. P. Veldkamp, and A. J. Verschoor, "Infeasibility in automated test assembly models: A comparison study of different methods," *J. Educational Measurement*, 2005.

[9] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[10] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, art. e1355, 2020.

[11] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Computers in Human Behavior*, vol. 89, pp. 98–110, 2018, doi: 10.1016/j.chb.2018.07.027.

[12] F. Wang *et al.*, "Neural cognitive diagnosis for intelligent education systems," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 4, 2020, pp. 6153–6161.

[13] C.-M. Chen, H.-M. Lee, and Y.-H. Chen, "Personalized e-learning system using item response theory," *Computers & Education*, vol. 44, no. 3, pp. 237–255, 2005.

[14] K. Wauters, P. Desmet, and W. Van den Noortgate, "Adaptive item-based learning environments based on the item response theory: Possibilities and challenges," *J. Computer Assisted Learning*, vol. 26, no. 6, pp. 549–562, 2010.

[15] G.-J. Hwang, B. M. T. Lin, and T.-L. Lin, "An effective approach for test-sheet composition with large-scale item banks," *Computers & Education*, vol. 46, no. 2, pp. 122–139, 2006.

[16] G.-J. Hwang, B. M. T. Lin, S.-H. Tseng, and T.-L. Lin, "On the development of a computer-assisted testing system with genetic test sheet-generating approach," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 35, no. 4, 2005.

[17] A. Ghosh and A. Lan, "BOBCAT: Bilevel optimization-based computerized adaptive testing," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2021.

[18] Y. Zhuang *et al.*, "Fully adaptive framework: Neural computerized adaptive testing for online education," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, no. 4, pp. 4734–4742, 2022, doi: 10.1609/aaai.v36i4.20399.

[19] A. C. M. Yang, B. Flanagan, and H. Ogata, "Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning," *Computers and Education: Artificial Intelligence*, vol. 3, art. 100104, 2022.

[20] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *J. American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.

[21] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical Science*, vol. 16, no. 2, pp. 101–133, 2001.

[22] B. Efron and C. Morris, "Stein's estimation rule and its competitors—an empirical Bayes approach," *J. American Statistical Association*, vol. 68, no. 341, pp. 117–130, 1973.

[23] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253–278, 1994.

[24] C. Piech *et al.*, "Deep knowledge tracing," in *Advances in Neural Information Processing Systems*, 2015, pp. 505–513.

[25] T. J. J. M. Theunissen, "Binary programming and test design," *Psychometrika*, vol. 50, no. 4, pp. 411–420, 1985.

[26] W. J. van der Linden, *Linear Models for Optimal Test Design*. New York, NY, USA: Springer, 2005.

[27] W. J. van der Linden and C. A. W. Glas, Eds., *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer, 2000.

[28] H. L. Roediger and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," *Psychological Science*, vol. 17, no. 3, pp. 249–255, 2006.

[29] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis," *Psychological Bulletin*, vol. 132, no. 3, pp. 354–380, 2006.

 This work is licensed under a Creative Commons Attribution 4.0 International License