

Toward Deployable Examination Paper Generation for Cross-Major Programming Assessment: A PRISMA-Guided Review of Multi-Constraint Optimisation and Hybrid GA-SA

Xu Lin

Sichuan Technology and Business University, ChengDu, 610000, China
INTI International University, Negeri Sembilan, 71800, Malaysia
ORCID iD: <https://orcid.org/0000-0002-2358-3013>
Email: 1062989925@qq.com

Marini Othman

INTI International University, Negeri Sembilan, 71800, Malaysia
ORCID iD: <https://orcid.org/0000-0001-7634-1686>
Email: marini.othman@newinti.edu.my

Tin Tin Ting*

Monash University Malaysia, Subang Jaya, 47500, Malaysia. UNITAR International University, Petaling Jaya, 47301, Malaysia
ORCID iD: <https://orcid.org/0000-0001-7634-1686>
Email: ting.tintin@monash.edu

Yuhang Wang

Sichuan Technology and Business University, ChengDu, 610000, China
ORCID iD: <https://orcid.org/0009-0009-3739-3788>
Email: wangyh12@alu.uestc.edu.cn

Recibido / Received: 09/07/2026
Aceptado / Accepted: 15/11/2026

Abstract: Automated Examination Paper Generation (EPG) has become a critical component of programming education, particularly in contexts where a common item bank is required to support multiple student cohorts characterised by varying difficulty levels, cognitive priorities, and skill-coverage needs. To establish a review-informed foundation for the practical design of deployable EPG systems in such environments, this study undertakes a PRISMA 2020-guided systematic literature review that conceptualises EPG as a multi-constraint combinatorial optimisation problem. By adopting a conservative evidence boundary together with a two-tier coding framework, the review integrates the ways in which earlier studies have operationalised structural requirements, difficulty levels, score allocation, skill coverage, Bloom's taxonomy-aligned cognitive domains, and efficiency-related performance measures. In parallel, it critically examines whether these studies provide deployment-relevant details, including item-bank metadata, transparency in constraint-handling procedures, trade-off decision rules, benchmark comparisons, and evidence relating to runtime performance and scalability. Across the reviewed literature, solution approaches extend from basic heuristics and exact optimisation models to metaheuristic methods, hybrid GA-SA architectures, and learning-augmented techniques. The synthesis indicates that hybrid GA-SA is consistently presented as a credible engineering framework for addressing dense, multi-objective EPG problems. Nevertheless, the existing body of literature remains limited by incomplete methodological reporting, insufficient machine-readable formalisation of pedagogical objectives, and a lack of evidence relevant to governance and implementation. Rather than asserting the universal superiority of any single algorithm, this review converts diverse findings into a deployability-oriented synthesis grounded in MCDM principles. Four major design implications are identified for programming-focused EPG systems: clearly

defined and auditable constraints, Bloom- and skill-aligned parameter specification, teacher-in-the-loop calibration mechanisms, and reproducible evaluation procedures. Collectively, these findings establish a defensible link between algorithmic selection and pedagogical application, while also providing a robust evidence base for future hybrid GA-SA development in cross-major programming assessment contexts.

Keywords: Examination Paper Generation (EPG), Systematic Literature Review (PRISMA), Constraint Handling, Hybrid Genetic Algorithm Simulated Annealing (GA-SA), Deployability (Runtime/Scalability Reporting), Multi-Criteria Decision Making (MCDM), Education.

1. Introduction

Global higher education is experiencing an accelerated phase of digital transformation, marked by deeper integration across student information systems, learning platforms, and digital assessment infrastructures (OECD, 2023). In China, this movement is further strengthened by policy initiatives focused on the governance and evaluation of digital and AI-enabled education, thereby intensifying the need for scalable and academically defensible assessment systems. Within such environments, automated assessment technologies are increasingly employed to minimise repetitive academic workload and maintain consistency as student enrolment expands (Chi, 2026; OECD, 2023). Nevertheless, for the purpose of this review, these broader developments are treated solely as contextual justification, while the evidential foundation remains strictly confined to studies retrieved through the PRISMA-guided search and screening process.

In contexts involving small item banks or comparatively flexible blueprints with limited constraints, straightforward assembly mechanisms may still remain feasible. However, these approaches tend to become progressively unstable and operationally inefficient as item-bank volume, blueprint precision, and constraint complexity increase. From an operations research standpoint, EPG may therefore be conceptualised as a multi-constraint optimisation problem in which items are drawn from an item bank to satisfy mandatory constraints while concurrently optimising one or more soft objectives (Van der Linden, 2005). Research on large-scale constrained test assembly similarly demonstrates that both infeasibility and instability become increasingly probable as the decision space widens and the number of constraints multiplies, thereby underscoring the necessity for systematic optimisation rather than ad hoc search strategies (Spaccapanico Proietti, Matteucci & Mignani, 2020).

This issue becomes particularly significant in the context of programming assessment. Computer science is not exclusively taught to students specialising in programming; it is also undertaken across other academic disciplines and by research students, whose learning models and competency expectations differ substantially (Gao et al., 2025; Riese & Bälter, 2022). As a result, a single examination structure is rarely appropriate for all programming-related cohorts, since learners from diverse academic backgrounds demonstrate varying levels of skill and domain knowledge. Within the evidence-centred design (ECD) framework, assessments are expected to align with the intended construct, task characteristics, scoring logic, and the inferential relationship between observable student performance and claims regarding knowledge and competence (Mislevy, Almond & Lukas, 2003).

Contemporary EPG research has evolved beyond basic paper assembly toward specification-driven optimisation frameworks. Recent studies increasingly model multiple objectives simultaneously, including difficulty balance, score distribution, content alignment, cognitive-level representation, and skill coverage, rather than relying on a single quality

criterion (Shang et al., 2023; Wang et al., 2024; Wu et al., 2020). At the same time, the algorithmic landscape has expanded considerably. Existing approaches now include baseline strategies, random search, and backtracking; exact yet tractable methods such as ILP, MIP, and CP; heuristic techniques including GA and SA; hybrid frameworks; and, more recently, learning-enhanced and reinforcement learning-based models (Shang et al., 2023; Wang et al., 2024; Wu et al., 2020). Among these, hybrid GA-SA has emerged as a particularly prominent approach due to its repeated recognition as an exploration-refinement framework that combines global search capability with local optimisation. This structure potentially reduces the risk of premature convergence while improving robustness under dense and conflicting constraints (Yang, 2015). Such characteristics are especially relevant where examination papers must simultaneously satisfy multiple pedagogical and operational requirements.

Algorithmic evaluation also raises critical issues concerning fairness, transparency, and explainability, particularly in relation to hidden biases embedded within automated systems (Baker & Hawn, 2022; Khosravi et al., 2022). Concurrently, pedagogical intent often remains insufficiently formalised. Dimensions such as Bloom-aligned cognitive levels, programming skill domains, and blueprint tolerance are frequently discussed only at a descriptive level rather than being translated into machine-readable tags, explicit constraints, or objective functions. Bloom-aligned parameterisation provides one viable mechanism for converting instructional intent into formal specifications suitable for automated generation (Krathwohl, 2002). Similarly, item banks and their associated metadata should not be viewed merely as implementation components; rather, they constitute a fundamental deployability infrastructure for EPG systems (Chituc et al., 2019).

Accordingly, this study adopts an MCDM lens to evaluate EPG methods across the dimensions of feasibility, efficiency, pedagogical alignment, and transparency (Triantaphyllou, 2000). In this way, the review seeks to establish a rigorous connection between algorithmic design and educational practice, while simultaneously laying the conceptual groundwork for subsequent discussion of hybrid GA-SA applications in cross-major programming assessment. Consequently, this review addresses the following research questions:

- RQ1: How have the objectives and constraint complexity in EPG research evolved over time?
- RQ2: Which solver families have been used for multiconstrained EPG in educational assessment, especially programming education, and what problem settings where they designed to address?
- RQ3: To what extent do existing EPG systems accommodate multiple cognitive levels and programming-skill dimensions, and how explicitly are these intentions parameterised?
- RQ4: What interdisciplinary frameworks can bridge algorithm development and teaching practice through teacher participation and Bloom-aligned learning goals?

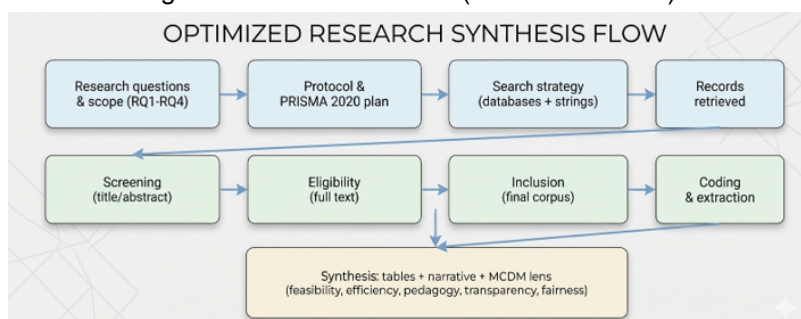
The remainder of this paper is structured as follows. Section 2 outlines the PRISMA-guided methodology and defines the scope boundaries of the review. Section 3 presents a synthesis of the literature concerning EPG formalisation, solver families, and existing reporting deficiencies. Section 4 interprets the findings through a deployability-oriented discussion grounded in an MCDM perspective, with specific emphasis on teacher-in-the-loop calibration and Bloom-aligned parameterisation. Finally, Section 5 summarises the review, discusses its limitations, and proposes directions for future research.

2. Methodology

2.1. PRISMA Methodology

This review adheres to the PRISMA 2020 statement and its accompanying explanatory guidance to report the processes of study identification, selection, eligibility assessment, and synthesis relevant to EPG (Page et al., 2021). In line with the evidence boundary established for this paper, both the synthesis and coding procedures are based solely on information that is explicitly reported and accessible within the reviewed texts. No details are assumed or inferred where such information is absent in the original reporting (Figure 1).

Figure 1: Review Workflow (PRISMA-Guided).



The review protocol defines the following components: (i) the research questions (RQ1–RQ4), (ii) the scope of EPG as the construction of examination papers from an item bank, (iii) the inclusion and exclusion criteria, and (iv) the table-based coding framework employed to enable an auditable cross-study synthesis. The boundary of coded evidence is established through Tables 1 and 2 and is strictly confined to EPG studies. Related literature concerning ATA, adaptive testing, and psychometric optimal test design is referenced solely to provide conceptual and methodological context and is therefore excluded from the coded frequency counts (Becker et al., 2021; Spaccapanico Proietti et al., 2020; Van der Linden, 2005, 2022).

2.2. Search Strategy and Information Sources

The search strategy was developed to ensure balanced coverage across both educational technology sources and optimisation/AI-focused publication venues. The databases consulted included Google Scholar, Clarivate, IEEE Xplore, Association for Computing Machinery Digital Library, and Springer Nature. The search period spanned from 2006-01-01 to 2026-02-28, with particular emphasis placed on the most recent five years to capture developments in learning-enhanced and deployability-oriented EPG. The search strings were constructed by combining three conceptual groups through the AND operator: (A) examination paper generation and related descriptors (for example, exam/test paper generation and automatic paper composition), (B) optimisation methods and solver families (such as genetic algorithm/GA, metaheuristic approaches, reinforcement learning/RL, and integer programming), and (C) constraints and objective dimensions

(including coverage, difficulty, score distribution, and skill/Bloom alignment). The precise syntax was customised for each database in accordance with its field tags, wildcard rules, and proximity operators. Final search exports were systematically archived for subsequent analysis and data extraction.

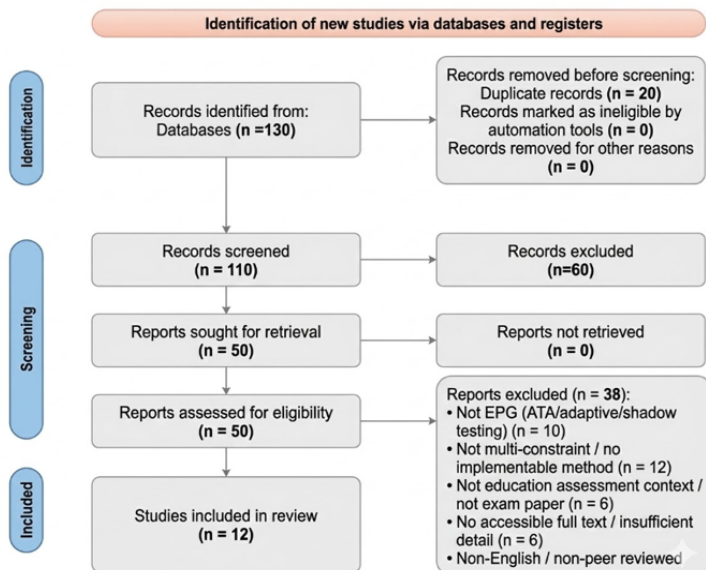
2.3. Eligibility Criteria

The inclusion criteria were defined as follows: (1) the study develops, evaluates, or implements an automated approach for assembling an examination paper from an item bank under specified constraints and/or objectives; (2) the study addresses constraints beyond merely trivial identification; (3) the study offers a primary methodological contribution, such as an algorithm, system architecture, or evaluation framework; and (4) the full text is available and provides sufficient methodological detail to assess the constraint-objective design. Even where the full text was accessible, a study could still be coded as incomplete or insufficient in Table 2 if key methodological elements, including metadata provenance, trade-off rules, runtime evidence, or baseline definitions, were reported inadequately. The exclusion criteria comprised the following: (1) studies concerned exclusively with ATA of parallel forms, adaptive shadow testing, or psychometric form-equating workflows; (2) papers in which EPG was mentioned only briefly without providing a usable methodological framework; and (3) publications that were either non-English or not peer reviewed.

2.4. Study Selection and Screening Process

All retrieved records underwent a two-stage screening process: first, title and abstract selection, followed by full-text eligibility assessment. Decisions, along with the corresponding reasons for exclusion, were recorded in a screening log to ensure full auditability.

Figures 2: PRISMA 2020 Flow Diagram for Study Selection (2006-01-01 to 2026-02-28).



The PRISMA flow counts were as follows: records identified from databases, N=130; duplicates removed, N=20; title and abstracts selected, N=110; title and abstracts excluded, N=60; full texts assessed for eligibility, N=50; full texts excluded, N=38; and studies included in the final review, N=12. Exclusion at the full-text stage was due to reasons such as the study not addressing EPG, failing to consider multi-constraint or implementable methods, not being situated within an examination context, providing insufficient methodological detail, or not meeting language or peer-review standards (Figures 2).

2.5. Data Extraction and Coding Scheme

A structured extraction form captured the following information: bibliographic metadata (year, publication venue); problem context (course or domain and paper specifications); item bank characteristics (reported size, metadata tags, and provenance); constraint-objective parameterisation; solver family and feasibility-enforcement mechanisms; and evaluation evidence, including feasibility, proxies for objectives or quality, runtime and scalability, baselines, and metric computability. To ensure auditable cross-study synthesis, Section 3 is organised into two interconnected coding levels derived from this extraction framework. Table 1 documents the explicit parameterisation of key instructional intent dimensions within constraints and objectives, while Table 2 records the minimum reporting requirements necessary for deployability-oriented interpretation and cross-study comparability.

2.6. Evaluation of Evidence Quality and Reporting

Given that EPG studies are conducted across diverse item banks and assessment contexts, this review does not assign a single composite quality or bias score. Instead, the evidential strength supporting deployment-oriented conclusions is assessed through a minimum reporting checklist (Table 2), which encompasses: (a) item bank size and metadata provenance; (b) transparency of constraints, including hard/soft designation and penalty or enforcement mechanisms; (c) specification of objectives and trade-offs, including weights, aggregation functions, and decision rules; (d) evidence on runtime, computational environment, and scalability; and (e) baselines and metric computability. Tables 1 and 2 fulfil distinct functions. A study may meaningfully parameterise pedagogical intent as captured in Table 1, yet still omit deployment-critical details such as metadata provenance, explicit trade-off rules, runtime conditions, or reproducible baselines, as recorded in Table 2.

2.7. Synthesis Approach

The synthesis is conducted across three levels. First, table-driven coding is employed to characterise how EPG is specified and which instructional intent dimensions are explicitly parameterised (Table 1). Second, solver families are compared with respect to their typical applications, feasibility-enforcement mechanisms, commonly reported evaluation evidence, and recurrent deployment-relevant gaps (RQ2). Third, the review findings are interpreted through a deployability-oriented lens, resulting in a decision framework and a minimum deployable bundle (MDB) that encompasses machine-readable intent specifications, reporting requirements, and audit-ready artefacts (RQ4).

3. Results

3.1. Overview of the Section

This section presents a synthesis of evidence from the PRISMA-guided systematic literature review on EPG, with a focus on multi-constraint combinatorial optimisation for scalable programming assessment. Section 3.1 formalises the EPG problem and introduces a constraint-objective taxonomy. Section 3.2 reviews principal solver families, including baselines, exact methods, metaheuristics, hybrid approaches such as GA-SA, and learner-aware methods. Section 3.3 consolidates evaluation practices and identifies reporting deficiencies that constrain deployability-oriented conclusions, encompassing concerns related to validity, transparency, and fairness. Section 3.4 subsequently translates these findings into a decision-oriented synthesis that underpins the MCDM perspective in Section 4 and informs a review-derived roadmap for teacher-in-the-loop calibration and Bloom-aligned parameterisation.

The mapping of research questions is as follows: Section 3.1 examines the evolution of goals and constraints over time, addressing the ‘as specified’ element of RQ3. Section 3.2 synthesises solver families alongside the problem contexts to which they are applied (RQ2). Section 3.3 evaluates the types of evidence studies provide beyond mere feasibility, including validity proxies, auditability, traceability, robustness, and fairness-related considerations. Section 3.4 integrates these insights into a deployability-oriented synthesis that supports the interdisciplinary bridge developed in Section 4 (RQ4). Within this review, the term EPG is employed as a general label for constraint-based examination paper assembly from an item bank. Closely related terms used in some studies, such as test form construction, are incorporated only when their problem formulations, solver selections, or evaluation practices offer direct relevance to deployable EPG (Huang et al., 2023).

3.2. Formalisation of the EPG Problem and Constraint-Objective Taxonomy

This section defines the constraint-objective taxonomy employed to code the EPG corpus consistently across temporal evolution (RQ1), solver families (RQ2), parameterised pedagogical intent (RQ3), and practice-oriented implications (RQ4). Across the reviewed literature, it is increasingly evident that producing a merely feasible examination paper is no longer adequate. The primary challenge lies in balancing feasibility, quality, and efficiency while adhering to increasingly detailed instructional specifications (Shang et al., 2023; Wang et al., 2024; Wu et al., 2020).

3.2.1. Formalisation: Decision Variables, Specifications, and Outputs

Definition

In the reviewed EPG studies, the primary task involves selecting a subset of items from a larger item bank to construct an examination document (Shang et al., 2023; Wang et al., 2024).

Multi-Constraint and Large Search Space

Accordingly, EPG constitutes a multi-constraint optimisation problem defined

over a large combinatorial search space, in which multiple rules must be satisfied simultaneously and several solution-quality objectives may coexist (Huang et al., 2023; Shang et al., 2023; Wang et al., 2024).

Decision variables and outputs. Binary selection variables are commonly used in EPG models, where an item is assigned a value of 1 if selected and 0 otherwise. A feasible assignment corresponds to a valid examination paper that meets structural specifications such as total score, number of items, item-type composition, and, where applicable, time constraints (Han, Gao & Zhai, 2023; Wang & Wang, 2022).

Co-Modelling of Quality and Efficiency

Recent studies additionally incorporate generative efficiency alongside examination quality, with quality captured through metrics such as knowledge-point coverage, cognitive level, and item difficulty (Wang et al., 2024).

3.2.2 Constraint Taxonomy: Hard Feasibility vs. Soft Preferences

Hard Versus Soft Constraints

Specifications in EPG may be implemented as hard feasibility constraints, which must be strictly satisfied, or as soft preferences expressed through deviation or penalty terms incorporated into the optimisation objective (Farahmand-Tabar & Sadrekarami, 2023).

Soft Constraints as Operational Targets

In contemporary EPG formulations, soft constraints are frequently operationalised as deviation measures, such as the distance from a target difficulty profile or score distribution, and are optimised directly rather than being applied solely as post hoc filters (Shang et al., 2023; Wang et al., 2024).

Educational Attributes Commonly Encoded in the Corpus

Across the reviewed studies, constraints and optimisation targets typically involve the following dimensions, which are subsequently coded in Table 1:

1. Knowledge/Content Coverage

Coverage targets are implemented either as hard constraints or as quality objectives, often expressed through coverage rates or deviations from the intended content balance (Wang et al., 2024).

2. Score Distribution

Recent formulations explicitly optimise score distribution alongside difficulty and skill coverage, extending earlier approaches that concentrated primarily on difficulty (Shang et al., 2023; Wu et al., 2020).

3. Skill Coverage

Learner-aware and data-driven EPG approaches increasingly incorporate skill coverage and the skill structure of items as explicit optimisation objectives (Shang et al., 2023).

4. Cognitive Level (Bloom-Aligned Demand)

When cognitive level is parameterised, studies specify the derivation of labels.

Current research in computing education and educational technology indicates that Bloom-aligned classification can be operationalised using machine-readable item tags, thereby making cognitive constraints and objectives more practicable (Gani et al., 2023; Kumar, Gulwani & Singh, 2025; Zhang et al., 2021).

5. Structure (Totals/Types/Time)

Total score, item composition, number of items, and time limits remain the foundational structural dimensions of EPG. Contemporary implementations explicitly encode these requirements within the assembly problem (Popescu, 2025; Wang & Wang, 2022).

3.2.3. Objective Functions: Beyond Feasibility to Multi-Objective Quality Optimisation

Quality-First Modelling

Recent EPG studies increasingly prioritise quality criteria, such as difficulty control, score distribution, and skill coverage, as primary optimisation targets rather than treating them solely as post-assembly evaluative measures (Shang et al., 2023).

Multi-Objective Specifications

Typical multi-objective EPG formulations simultaneously optimise difficulty, score distribution, and skill coverage to ensure that the generated examination paper aligns more closely with the intended assessment objectives (Shang et al., 2023).

Quality-Efficiency Coupling

Some studies explicitly integrate quality objectives with deployment-oriented efficiency requirements, optimising knowledge-point coverage, cognitive level, question difficulty, and generation efficiency concurrently (Wang et al., 2024).

Solution Choice as an Implementation Strategy, Not Proof of Superiority

GA-based methods are frequently employed as practical strategies for navigating large, constrained search spaces and enhancing paper quality through customised operators or hybridisation (Cui, Zhou & Huang, 2023; Wang et al., 2024). Learning-enhanced approaches additionally explore multi-objective optimisation via reinforcement learning (Shang et al., 2023). More generally, this approach aligns with metaheuristic design principles for realistic combinatorial optimisation contexts (Osaba et al., 2021).

Signal for RQ1

Collectively, the literature indicates a potential shift from earlier difficulty-centric formulations toward richer objective bundles that incorporate score distribution and skill coverage. However, the strength of this temporal inference is constrained by uneven reporting practices across studies (Shang et al., 2023).

3.2.4. Reporting and Specification Gaps: Minimum Checklist for Auditability and Deployability

Minimum Reporting Checklist

To facilitate cross-study comparison and support informed engineering judgements,

this review proposes a minimum reporting checklist for selected EPG studies. The checklist encompasses:

- (a) item-bank scale and metadata provenance, including the source of coverage, difficulty, skill, and cognitive labels;
- (b) constraint transparency, specifying hard/soft designation and deviation or penalty definitions;
- (c) objective and trade-off specification, including aggregation methods, weights, and decision rules;
- (d) runtime and scalability evidence, detailing hardware and time-cost reporting; and
- (e) baselines and metric computability, with clearly defined baselines and reproducible metric calculations (Shang et al., 2023; Wang et al., 2024; Wu et al., 2020).

This checklist is grounded in broader principles of reproducibility and auditability in optimisation and metaheuristic research. Incomplete reporting of experimental settings, baseline comparisons, runtime conditions, or measurable evaluation criteria hinders independent verification and undermines deployment-oriented interpretation (Ahn et al., 2022; Da Ros et al., 2026; Herring, Kirley & Yao, 2022). Comparable considerations are emphasised in AI auditability and educational AI governance research, which highlights the necessity of transparency, documentation, explanation, and scrutiny (Baker & Hawn, 2022; Khosravi et al., 2022; Li & Goel, 2025).

3.2.5. Summary

In summary, the review operationalises auditability and deployability using a checklist that encompasses: (1) data and metadata provenance; (2) constraint specification and handling; (3) objective and trade-off rules; (4) baselines and metric computability; and (5) runtime and scalability evidence. The coded distribution of these dimensions across the included EPG studies is presented in the table-driven cross-study synthesis below. Table 1 Constraint-Objective Taxonomy Coding for the EPG Corpus (N=12).

Table 1: Constraint-Objective Taxonomy Coding (/ /, EPG, N=12).

Study	Structural Totals/Types/Time	Coverage	Difficulty	Score Distribution	Skill Coverage	Cognitive/ Bloom	Efficiency as Objective
Veldkamp and Paap (2017)	✓	✓	✓	—	—	—	△
Wang and Wang (2022)	✓	✓	✓	✓	△	✓	△
Huang et al. (2023)	△	△	—	—	—	—	—
Cui et al. (2023)	△	—	△	—	—	—	△
Han et al. (2023)	✓	✓	✓	△	—	—	△
Shang et al. (2023)	△	—	✓	✓	✓	—	△
Popescu (2025)	△	△	—	—	—	—	△
Wang et al. (2024)	△	✓	✓	—	—	✓	✓
Popescu (2025)	✓	△	✓	△	—	—	△
Popescu (2025)	✓	✓	✓	—	—	—	△
Popescu (2025)	△	—	✓	✓	△	—	△
Wang et al. (2024)	△	—	△	—	△	—	—

Note: ✓ = explicitly operationalised with implementable detail; △ = mentioned but not operationally specified; — = not reported or unclear in the accessible text. Coding is based solely on explicit reports, and no capability is inferred.

3.3. Solver Families for EPG (How is it solved?)

Building on the formalisation and constraint-objective taxonomy presented in Section 3.1, this section synthesises the solver families employed in multi-constraint EPG using a deployability-oriented template. For each solver family, the review examines: (i) typical application settings, including item bank scale, constraint density, and objective structure; (ii) strategies for enforcing or recovering feasibility; (iii) commonly reported forms of evaluation evidence; and (iv) deployment-relevant information that is frequently omitted, as assessed against the checklist introduced in Section 3.1.4.

Scope Note

In line with the defined review boundaries, ATA pipelines for parallel forms, adaptive shadow testing, and psychometric form-equating workflows are excluded from the evidence base used to characterise solver families in this section.

3.3.1. Exact Formulations (ILP/MIP/CP)

Exact Methods

When to Use Them

Exact formulations, including integer linear programming (ILP), mixed integer programming (MIP), and constraint programming (CP), are most appropriate when the item bank is relatively small, constraints are tractable, or a formally defined and auditable baseline is required.

Feasibility Enforcement

Feasibility is maintained through explicit constraints, while trade-offs are encoded via formal objective structures such as weighted sums or lexicographic priorities. When fully documented, these formulations provide relatively high auditability.

Commonly Reported Evidence

Studies employing exact methods typically report whether solutions are optimal, feasible, or infeasible, alongside objective values and partial information on constraint satisfaction. Runtime and scalability, however, are highly sensitive to item bank size and constraint density and therefore require explicit reporting.

Deployment-Relevant Gaps

Even with an explicit mathematical formulation, the deployability of exact methods depends critically on the provenance of item-bank metadata and the computability of evaluation metrics. Without these elements, precise comparisons between real-world EPG systems are severely constrained.

3.3.2. Baselines and Rule-Based Heuristics

These methods are typically employed in small to moderate item banks with relatively few constraints, where test assembly involves strict optimisation. Techniques commonly used include rule screening, stepwise greedy checks, random sampling, and backtracking search (Huang et al., 2023). Feasibility is maintained through local

validation strategies such as skipping, swapping, resampling, or backtracking. These approaches function adequately when constraints are less stringent or become denser, although their effectiveness is limited by the absence of a comprehensive global view of the solution space. Baseline studies generally report whether solutions are feasible and provide approximate quality measures. However, detailed analyses of runtime scaling or target-level comparisons are rarely included.

Deployment-Relevant Gaps

Reproducibility is undermined when reporting is incomplete, particularly regarding filters, stopping rules, item-bank scale, metadata provenance, and metric computation procedures (Ahn et al., 2022; Da Ros et al., 2026; Herring et al., 2022).

3.3.3. Metaheuristics (GA and Related Stochastic Search)

Metaheuristics are primarily applied to large item banks and complex optimisation problems with numerous constraints, with GA-based EPG methods commonly used to enhance search quality under these conditions (Cui et al., 2023; Popescu, 2025).

Feasibility Enforcement: Common methods include penalties, repair operators, and feasibility-conserving approaches that maintain feasible solutions. This aligns with the broader shift from post hoc checking toward direct optimisation of specification satisfaction.

Commonly Reported Evidence: Metaheuristic papers more often report objective value surrogates, such as coverage deviation, difficulty deviation, and conformity to target distributions, and sometimes include runtime information (Cui et al., 2023; Popescu, 2025). However, important details for reproduction, such as parameter settings, stopping rules, and sensitivity to hyperparameters, are unevenly reported.

Deployment-Relevant Gaps: Traceability artefacts, such as constraint-satisfaction reports, decision rationales, replicable scripts, and configuration files, are less consistently reported than headline objective values, making deployment-orientated comparison difficult.

3.3.4. Hybrid and Multistage Metaheuristics (Hybrid GA Templates and GA-SA Rationale)

When They Are Used: Hybridisation is generally employed when constraints are strict or when multiple objectives need to be balanced concurrently. The EPG literature presents hybrid GA variants as practical methods for constrained paper generation (Wang & Wang, 2022). Some studies also integrate structured subroutines, such as dynamic programming-guided genetic search, particularly when repeated subassembly patterns are significant (Wang et al., 2024). Combined GA-SA approaches leverage global exploration alongside constraint handling and local refinement within schedules and rules, which are described clearly (Yang, 2015).

Commonly Reported Evidence

Hybrid studies typically focus on feasibility and solution-quality proxies and may report runtime or efficiency improvements when efficiency is treated as an explicit objective (Wang et al., 2024). However, sensitivity to algorithm parameters is rarely documented.

Deployment-Relevant Gaps

Hybrid methods introduce additional hyperparameters and stage-wise logic, making them particularly dependent on comprehensive reporting of algorithm configuration and experimental protocols to support defensible reproduction and deployment claims (Ahn et al., 2022; Da Ros et al., 2026; Herring et al., 2022).

3.3.5. Learner-Aware / Data-Driven and RL-Based Approaches

When They Are Used: Learning-enhanced EPG is particularly relevant when the objective bundle is complex or when learner-specific information is leveraged to guide paper assembly. Recent studies utilise reinforcement learning to jointly optimise difficulty, score distribution, and skill coverage, aiming to align more closely with assessment intent (Shang et al., 2023). Personalisation-oriented variants additionally incorporate knowledge-state modelling to tailor papers for individual students.

Feasibility Enforcement: Reinforcement learning-guided EPG maintains feasibility through rewards, penalties, repair operators, or feasibility-conserving actions.

Commonly Reported Evidence: These studies typically report improvements in multi-objective outcomes and, less consistently, efficiency-related metrics (Shang et al., 2023). However, reliance on training data, label quality, and policy transparency introduces additional deployment risk.

Deployment-Relevant Gaps: Learning-based approaches demand strong auditability, traceability, and transparent AI pipelines to ensure defensible deployment (Baker & Hawn, 2022; Khosravi et al., 2022; Li & Goel, 2025).

3.3.6. Summary (Links to RQ2 and Section 3.1.4)

Overall, Section 3.2 synthesises five practically relevant solver families for interpreting multiconstraint EPG: (i) exact formulations; (ii) baseline or rule-based heuristics; (iii) metaheuristics; (iv) hybrid or multistage metaheuristics; and (v) learner-aware, data-driven, and RL-based approaches. These solver families vary not only in their optimisation mechanisms and approaches to maintaining feasibility but also in the extent to which they document deployment-critical evidence. Therefore, the comparisons presented in this review are based on the constraint-objective taxonomy from Section 3.1 and the reporting standards outlined in Section 3.1.4, rather than any assumption of inherent superiority.

3.4. Evaluation, Reporting Practices, and Evidence Gaps

This section examines PRISMA-compliant EPG studies with a focus on reproducibility, auditability, and evidence pertinent to practical deployment.

3.4.1. Table-Driven Evidence Coding and Scope Boundary (EPG)

To maintain auditable cross-study synthesis, the review adopts a table-driven coding strategy. Table 1 operationalises the constraint-objective taxonomy from Section 3.1 by indicating whether key intent dimensions are explicitly parameterised. Table 2 codes the minimum reporting checklist necessary for deployment-orientated interpretation and cross-study comparison, covering: (a) item-bank scale and metadata

provenance; (b) constraint transparency; (c) objective and trade-off rules; (d) runtime, environment, and scalability evidence; and (e) baselines and metric computability. The coded evidence base is limited to EPG studies. Related literature on ATA, adaptive testing, and optimal psychometric test design is cited only for conceptual framing and does not contribute to coded counts or comparative claims (Becker et al., 2021; Spaccapanico Proietti et al., 2020; Van der Linden, 2005, 2022).

Table 2: Coding of the Minimum Reporting Checklist (/ , EPG, N=12).

Study	Bank Scale & Metadata	Constraint Transparency	Objective & Trade-Off Rule	Runtime & Env/ Scalability	Baselines & Metric Computability
Xu et al. (2024)	△	✓	✓	△	△
Wang and Wang (2022)	△	△	△	△	—
Huang et al. (2023)	—	—	—	—	—
Cui et al. (2023)	△	—	—	△	✓
Han et al. (2023)	△	✓	△	△	✓
Shang et al. (2023)	✓	△	△	△	△
Xu et al. (2024)	△	△	—	△	△
Wang et al. (2024)	✓	✓	✓	✓	✓
Xu et al. (2024)	✓	✓	✓	△	✓
Popescu (2025)	△	✓	✓	△	✓
Xu et al. (2024)	✓	△	△	△	△
Popescu (2025)	△	—	—	—	✓

Note: The dimensions are: (a) item-bank scale and metadata provenance; (b) constraint transparency, including hard versus soft designation and penalty / device definition; (c) objective and trade-off rules, including weights, aggregation and decision criteria; (d) run-time, environment and scalability evidence; and (e) baselines and metric computability. Coding is based on explicit reporting only; no capability is inferred.

3.4.2. What is Commonly Reported: Feasibility and Blueprint Alignment Proxies

In the EPG corpus, the most consistently reported outcomes are feasibility achievement and solution-quality indicators. Many studies report whether feasible papers can be generated and provide optimisation results or blueprint-alignment proxies, such as knowledge coverage and difficulty. In learner-aware approaches, these quality proxies become more nuanced and may also encompass score distribution and skill coverage (Shang et al., 2023; Wu et al., 2020). Since not all studies parameterise identical pedagogical intentions, improvements in a subset of proxies do not necessarily ensure broader pedagogical or deployable alignment, highlighting the need for Table 2.

3.4.3. Deployment-Critical Gaps: Trade-Offs, Reproducibility, Runtime/ Scalability, and Baselines

Compared with feasibility and headline objective values, deployment-critical reporting is less consistently addressed. Table 2 highlights several recurrent shortcomings: (1) explicit trade-off rules in multi-objective contexts are often absent, limiting fair comparison under similar preference structures (Triantaphyllou, 2000); (2) runtime, environment, and scalability evidence is frequently insufficient, despite being essential for evaluating practical operability in expanding digital assessment ecosystems (OECD, 2023); (3) baselines and metric computability are not always transparently reported; and (4) details of constraint handling, including penalty schedules, violation

measures, and repair procedures, are often under-specified. These issues reflect broader concerns in optimisation reproducibility research (Ahn et al., 2022; Da Ros et al., 2026; Herring et al., 2022) and are consistent with best-practice guidance for applied metaheuristics (Das et al., 2025; Gendreau & Potvin, 2019; Osaba et al., 2021).

3.4.4. Governance-Facing Gaps: Auditability, Explainability, and Bias/Fairness Risk

Beyond reproducibility, deployable EPG demands evidence that supports governance and accountability. Research on algorithmic bias in education highlights that automated decisions can perpetuate existing inequalities, necessitating supervision mechanisms that enable stakeholders to interpret and challenge system behaviour (Baker & Hawn, 2022). Similarly, work on explainable AI emphasises the need for intelligible system behaviour when educational decisions are influenced by automated processes (Khosravi et al., 2022). From an auditability perspective, deployable systems should provide documentation and engineering artefacts that facilitate effective review when auditing is required (Li & Goel, 2025). Nevertheless, within the reviewed EPG corpus, explicit reporting on auditability artefacts—such as constraint satisfaction summaries, reasoning traces, configuration versioning, or robustness and fairness diagnostics—remains scarce.

3.4.5. Evidence Hook for RQ3: Intent Parameterisation of Cognitive Level and Programming Skills

RQ3 examines the extent to which EPG systems accommodate diverse cognitive levels and programming skills, and how explicitly these intentions are parameterised. Within the evidence boundary of this review, robust evidence requires that such pedagogical goals be encoded as machine-interpretable specifications that are incorporated into the optimisation process via constraints or objectives. A study is coded positively only when it explicitly operationalises cognitive/Bloom levels or skill coverage using item tags with clearly stated provenance and integrates these into blueprint quotas, hard/soft constraints, or objective/penalty functions. Narrative mentions lacking operational integration are coded as weak or absent evidence. This standard aligns with both the foundational and revised Bloom frameworks (Krathwohl, 2002) and is supported by research demonstrating that automated methods can classify questions and learning outcomes by Bloom level, thus enabling machine-readable parameterisation (Gani et al., 2023; Kumar et al., 2025; Zhang et al., 2021). In programming education, studies on learner heterogeneity further reinforce the need for skill- and cohort-aware blueprinting and reporting (Gao et al., 2025; Riese & Bälter, 2022).

3.4.6. Implications for Synthesis Strength (Path to Section 3.4 and Section 4)

Taken together, Tables 1 and 2 clarify why this section does not present a final ranking or declare a “best” approach. Table 1 documents the variation in pedagogical and operational intentions that were explicitly parameterised, while Table 2 records the completeness of deployment-relevant evidence reporting. Consequently, Sections 3.3 and 3.4 adopt a synthesis constrained by the available evidence, refraining from

inferring capabilities beyond what is explicitly reported. This cautious stance underpins the decision-orientated translation presented in the following section.

3.5. Decision-Orientated Synthesis for Deployable EPG (So What?)

Section 3.4 synthesises the insights from Tables 1 and 2 without establishing a single ranking. Its purpose is to serve as a conceptual bridge to Section 4 by delineating the limits of what can be concluded from the available evidence and by translating the review findings into decision-relevant guidance addressing RQ1–RQ4.

3.5.1. Why a Single-Metric Ranking is Inadequate (Motivation for MCDM)

Deploying EPG requires balancing trade-offs among feasibility under strict constraints, solution quality, efficiency and scalability, pedagogical validity, transparency, traceability, robustness, and fairness-related risk. Given the substantial variation in parameterisation breadth (Table 1) and reporting completeness (Table 2) across studies, a single-metric ranking would inappropriately treat heterogeneous evidence as directly comparable. A MCDM perspective is therefore more suitable, as it explicitly represents trade-offs and translates diverse evidence into a defensible selection framework rather than asserting a universally optimal algorithm (Triantaphyllou, 2000). This approach also aligns with governance-oriented concerns in educational AI, which prioritise transparency, traceability, and explainability alongside technical performance (Baker & Hawn, 2022; Khosravi et al., 2022; Li & Goel, 2025).

3.5.2. Review-Derived Roadmap for RQ4: Teacher-in-the-Loop Calibration + Bloom-Aligned Targets

RQ4 indicates that the teacher-in-the-loop model connects optimisation algorithms with Bloom-aligned pedagogical objectives (Krathwohl, 2002). This approach also aligns with item bank engineering, as successful deployment relies on metadata quality, governance protocols, and operational constraints (Chituc et al., 2019).

Evidentiary Hook for Section 4

A deployable interdisciplinary bridge necessitates auditable artefacts that link pedagogical intent (blueprints and Bloom-aligned objectives), algorithmic execution (solver configurations and diagnostics), and evaluation outputs (constraint-satisfaction reports, deviation summaries, and reasoning logs). This approach aligns with evidence-centred design (Mislevy et al., 2003), which highlights the importance of an explicit claim-evidence-specification relationship. Accordingly, Section 4 translates the findings from this review into decision-orientated implications for deployment.

4. Discussion and Engineering Implications

4.1. Section Overview and Evidence Boundary

This section interprets the PRISMA findings with caution and emphasises deployability. In line with Table 2, the discussion makes only conservative claims and does not propose a definitive ranking of algorithms.

4.2. Direct Answers to RQ1–RQ4 (Conservative Summary)

RQ1: Objectives and Complimentary Constraint Over Time

The reviewed studies indicate a progression from relatively simple feasibility-focused assembly formulations toward more complex objective bundles that incorporate coverage, score distributions, skill coverage, and multiple trade-offs under denser constraints. Nonetheless, conclusions regarding temporal evolution are tentative, as many studies lack consistent reporting of constraint definitions, penalty structures, and baseline configurations.

RQ2: Which solver families are used and for what problem settings?

Section 3 synthesises five practically relevant solver families: (i) simple baseline or rule-based methods; (ii) exact formulations, including ILP, MIP, and CP, suitable when tractable; (iii) metaheuristics designed for large combinatorial search spaces; (iv) hybrid metaheuristics, such as GA combined with local refinement methods like SA, for denser multi-constraint problems; and (v) learning-enhanced or RL-guided approaches, particularly when learner data or richer objective bundles are incorporated. Solver selection should therefore be guided by alignment with the problem context rather than framed as a winner-takes-all decision.

RQ3: How are cognitive levels and programming skills accommodated?

Support for multiple cognitive levels and skill dimensions is most robust when these pedagogical intentions are encoded as machine-readable item tags and operationalised through explicit quotas, constraints, or penalty functions. Several studies, however, reference cognitive levels or skills only descriptively, without providing clear information on tag provenance, validation, or actual incorporation into the optimisation model.

RQ4: What interdisciplinary bridge links algorithm development and pedagogical practice?

A deployable implementation necessitates a teacher-in-the-loop workflow, integrating auditable optimisation procedures with governance-oriented design.

4.3. Cross-RQ Synthesis: Key Takeaways (C1–C5)

The discussion can be summarised into five key takeaways:

- C1. EPG should be viewed as a multi-objective, multi-constraint combinatorial optimisation problem. Deployability is enhanced when trade-offs among objectives are explicitly specified rather than obscured by a single aggregate score.
- C2. Exact methods offer auditable baselines and well-defined problem boundaries when tractable, while metaheuristics and hybrid approaches are more practical for large-scale or dense-constraint settings. These latter methods demand stronger and more transparent reporting to support reproducible deployment.
- C3. Pedagogical intent, including Bloom level and skill dimensions, is operationally meaningful only when encoded via explicit item tags that are directly linked to constraints, quotas, or penalty structures within the optimisation model.
- C4. Governance-oriented considerations—such as auditability, transparency,

explainability, and fairness risk management—are critical for educational deployment but are underreported in the existing EPG literature.

C5. Given that the strength of evidence depends on reporting completeness, deployment recommendations should be framed as a decision procedure or checklist rather than as a prescriptive ranking or classification of methods.

4.4. Decision Procedure for Deployment: An MCDM-Orientated Selection Guide

Building on C1–C5, this review proposes a review-informed, MCDM-orientated deployment procedure in which algorithm selection is explicit and linked to auditable evidence artefacts.

Step 1: Specify the Intent

Define blueprint targets, Bloom and skill dimensions, hard and soft constraints, and the tag schema along with its provenance.

Step 2: Choose a Solver Family

Select among exact, heuristic, hybrid, and learning-enhanced approaches based on: (a) item bank scale; (b) constraint density; (c) required transparency and auditability; and (d) available time and compute budget.

Step 3: Pre-Register Evaluation

Define feasibility metrics, objective metrics, baseline comparators, and runtime/ scalability criteria prior to experimentation rather than selecting them post hoc.

Step 4: Produce Deployability Artefacts

Publish definitions of constraints and penalties, solver configuration, random seeds where relevant, and a constraint-satisfaction report for generated papers.

Step 5: Conduct Governance Checks

Assess potential bias and fairness concerns related to item metadata or tagging, and explicitly state how such risks are to be managed, whether via constraints, post hoc review, or additional auditing procedures.

4.5. Interpreting Hybrid GA-SA in Light of the Review Evidence

The review highlights a recurring rationale for hybrid GA-style methods, including GA-SA variants, namely the combination of broad exploration with local refinement. This rationale is sensible in dense, multi-constraint settings, but the review does not position GA-SA as the superior approach. At minimum, reporting should include:

- (a) stage-wise logic, specifying GA operators, SA neighbourhood structure, cooling schedule, and termination rules;
- (b) explicit numerical details on constraint handling, including hard constraints, penalty terms, and repair operators;
- (c) robustness evidence, covering seed sensitivity, hyperparameter sensitivity, failure rates, and distributions of feasibility attainment; and
- (d) baseline comparisons under comparable item-bank, blueprint, and run-time conditions.

Under this perspective, hybrid GA-SA is best regarded as an engineering template whose plausibility relies on evidence quality and completeness of reporting, rather than as a universally optimal method.

4.6. Teacher-in-the-Loop and Bloom-Aligned Parameterisation as Deployability Infrastructure

Consistent with the evidence gaps identified in Section 3, teacher involvement should be operationalised as an auditable process rather than treated as a generic recommendation. A plausible minimal deployable design would include:

- (a) A versioned blueprint and tag taxonomy for Bloom levels and programming skills, accompanied by inter-rater or quality checks where feasible.
- (b) Iterative calibration rounds, allowing instructors to adjust weights, limits, or tolerances after reviewing sampled forms and constraint reports.
- (c) Explainability outputs targeted at stakeholders, illustrating why each generated paper satisfies the blueprint and where trade-offs were applied.
- (d) Periodic fairness and bias audits when relevant metadata may correlate with subgroup performance.

Viewed in this way, teacher-in-the-loop calibration and Bloom-aligned parameterisation should be regarded as essential components of a trustworthy deployment workflow rather than optional enhancements.

4.7. Limitations and Threats to Validity

The central shortcomings highlighted in Section 3 persist into this discussion. These include heterogeneous problem descriptions, inconsistent reporting of solver configurations and outcomes, sparse evidence on runtime and scalability, and limited governance-focused reporting on auditability and fairness. Consequently, the implications presented here should be interpreted as cautious guidance for method selection and deployment design, rather than as a prescriptive ordering of algorithms.

4.8. Section Summary

Section 4 translates the findings of Section 3 into a deployability-oriented synthesis. It provides conservative responses to RQ1–RQ4, consolidates the findings into five focused implications (C1–C5), and proposes an MCDM-style selection process. Within this context, hybrid GA-SA is treated as an engineering template requiring robust reproducibility and governance-facing evidence, while teacher-in-the-loop calibration and Bloom-aligned parameterisation are positioned as central components of a reliable EPG deployment workflow.

5. Conclusion and Future Directions

In accordance with the PRISMA 2020 workflow, this review synthesised evidence on automated EPG as a feasibility-gated, multi-objective, multi-constraint combinatorial optimisation problem. Aligned with the evidence boundaries defined in Section 3, the conclusions remain deliberately conservative, drawing only on what is explicitly

reported in the literature, including problem specifications, constraint configurations, solver characteristics, feasibility outcomes, and evaluation artefacts.

5.1. Limitations

The literature is challenging to compare because of variations in problem formulations, reporting thoroughness, and evaluation designs. Consequently, it is not possible to definitively identify any single algorithm family as superior. Critical deployment-related evidence—such as instance generators, constraint tightness, solver configurations, runtime budgets, traceability artefacts, and validated pedagogical labels—is often missing. Moreover, local tagging practices are commonly employed to assign educational intent labels, particularly for Bloom levels and programming skills, but these are frequently supported by limited validation data, which constrains the generalisability and transferability across courses and institutions.

5.2. Future Directions

Future research should emphasise interoperable evidence and governance-ready evaluation over isolated proof-of-concept demonstrations.

(a) Benchmarking and Controlled Instance Generation

Construct benchmark suites that systematically vary item-bank characteristics, constraint density or tightness, and intent profiles (e.g., skills × cognition), enabling rigorous assessment of sensitivity and robustness.

(b) Standardised Reporting Aligned with the MDB

Adopt reporting templates that capture specifications, solver configurations, baseline definitions, and teacher-in-the-loop evaluation designs, particularly in authentic classroom contexts spanning both computer science and non-computer science cohorts.

(c) Integrated Validity and Risk Assessment

Combine blueprint-alignment proxies with robustness testing, and where appropriate, implement fairness-risk audits in forms suitable for practical educational deployment.

(d) Stronger Reproducibility Culture

Encourage the use of explicit baselines, ablation studies, and open implementations or data sharing wherever feasible to address the current deficit in reproducible evidence.

5.3. Closing Remarks

Overall, the review highlights considerable methodological heterogeneity across EPG studies. Progress toward practical deployment is less about identifying a single superior algorithm and more about clearly defining, managing, and evaluating the relevant trade-offs. By linking optimisation strategies with pedagogical objectives, structured reporting, and governance considerations, the review provides guidance for defensible method selection and more dependable deployment of EPG in programming assessment across diverse disciplines.

Declarations

Funding

The cost of article processing is supported by INTI International University and Monash University Malaysia.

Conflict of Interest/Competing Interests

The authors have no competing interests to declare that they are relevant to the content of this article.

Ethical Approval

Not applicable.

Availability of Data and Material

Not applicable.

Reference

- Ahn, K., Jain, P., Ji, Z., Kale, S., Netrapalli, P. & Shamir, G. I. (2022). Reproducibility in Optimization: Theoretical Framework and Limits. *Advances in Neural Information Processing Systems*, 35, pp. 18022–18033. doi: <https://doi.org/10.48550/arXiv.2202.04598>
- Baker, R. S. & Hawn, A. (2022). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 32(4), pp. 1052–1092. doi: <https://doi.org/10.1007/s40593-021-00285-9>
- Becker, B., Debeer, D., Sachse, K. A. & Weirich, S. (2021). Automated Test Assembly in R: The eatATA Package. *Psych*, 3(2), pp. 96–112. doi: <https://doi.org/10.3390/psych3020010>
- Chi, W. J. (2026). Designing a pilot model for fully digitalised internal examinations in urban secondary schools: A sustainable transition framework. *Front. Educ.* 11:1684055. <https://doi.org/10.3389/educ.2026.1684055>
- Chituc, C.-M., Herrmann, M., Schiffner, D. & Rittberger, M. (2019). Towards the Design and Deployment of an Item Bank: An Analysis of the Requirements Elicited. In M. A. Herzog, Z. Kubincová, P. Han, & M. Temperini (Eds.), *Advances in Web-Based Learning – ICWL 2019* (pp. 155–162). Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-35758-0_15
- Cui, J., Zhou, Y. & Huang, G. (2023). A test paper generation algorithm based on diseased enhanced genetic algorithm. *Heliyon*, 9(6), pp. e17187. doi: <https://doi.org/10.1016/j.heliyon.2023.e17187>
- Da Ros, F., Začiragić, T., Plaat, A., Bäck, T. & van Stein, N. (2026). Assessing Reproducibility in Evolutionary Computation: A Case Study using Human- and LLM-based Assessment. *arXiv preprint arXiv:2602.07059*. doi: <https://doi.org/10.48550/arXiv.2602.07059>
- Das, A., Rai, R., Sasmal, B., Dhal, K. G., Khurma, R. A. & Saha, R. (2025). Metaheuristic Algorithms Since 2020: Development, Taxonomy, Analysis, and Applications. *Archives of Computational Methods in Engineering*. doi: <https://doi.org/10.1007/s11831-025-10408-3>
- Farahmand-Tabar, S. & Sadrekarimi, N. (2023). Overcoming Constraints: The Critical Role

- of Penalty Functions as Constraint-Handling Methods in Structural Optimization. In A. J. Kulkarni & A. H. Gandomi (Eds.), *Handbook of Formal Optimization* (pp. 1–26). Springer Nature Singapore. doi: https://doi.org/10.1007/978-981-19-8851-6_40-1
- Gani, M. O., Ayyasamy, R. K., Sangodiah, A. & Fui, Y. T. (2023). Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Education and Information Technologies*, 28(12), pp. 15893–15914. doi: <https://doi.org/10.1007/s10639-023-11842-1>
- Gao, Z., Yan, H., Liu, J., Zhang, X., Lin, Y., Zhang, Y., et al. (2025). Tracing distinct learning trajectories in introductory programming course: a sequence analysis of score, engagement, and code metrics for novice computer science vs. math cohorts. *International Journal of STEM Education*, 12(1), pp. 27. doi: <https://doi.org/10.1186/s40594-025-00546-2>
- Gendreau, M. & Potvin, J.-Y. (2019). *Handbook of Metaheuristics*. Springer. doi: <https://doi.org/10.1007/978-3-319-91086-4>
- Han, L., Gao, H. & Zhai, R. (2023). Modelling and Simulation of Intelligent English Paper Generating Based on SSA-GA. *Mathematical Problems in Engineering*, 2023(1), pp. 2277185. doi: <https://doi.org/10.1155/2023/2277185>
- Herring, D., Kirley, M. & Yao, X. (2022). Reproducibility and baseline reporting for dynamic multi-objective benchmark problems. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 529–537). Association for Computing Machinery. doi: <https://doi.org/10.1145/3512290.3528791>
- Huang, F., Wei, M., Su, L. & Guan, D. (2023). Design of a Dynamic Probability-Based Automatic Test Paper Generation System. In Z. Hu, I. Dychka, & M. He (Eds.), *Advances in Computer Science for Engineering and Education VI* (pp. 85–94). Springer Nature Switzerland. doi: https://doi.org/10.1007/978-3-031-36118-0_8
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., et al. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, 3, pp. 100074. doi: <https://doi.org/10.1016/j.caeai.2022.100074>
- Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4), pp. 212–218. doi: https://doi.org/10.1207/s15430421tip4104_2
- Kumar, R., Gulwani, D. & Singh, S. (2025). Automated Analysis of Learning Outcomes and Exam Questions Based on Bloom's Taxonomy. *arXiv preprint arXiv:2511.10903*. doi: <https://doi.org/10.48550/arXiv.2511.10903>
- Li, Y. & Goel, S. (2025). Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability. *Information Systems Frontiers*, 27(3), pp. 1121–1151. doi: <https://doi.org/10.1007/s10796-024-10508-8>
- Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). A Brief Introduction to Evidence-Centered Design. *ETS Research Report Series*, 2003(1), pp. i–29. doi: <https://doi.org/10.1002/j.2333-8504.2003.tb01908.x>
- OECD. (2023). *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*. OECD Publishing. doi: <https://doi.org/10.1787/c74f03de-en>

- Osaba, E., Villar-Rodriguez, E., Del Ser, J., Nebro, A. J., Molina, D., LaTorre, A., et al. (2021). A Tutorial On the design, experimentation and application of metaheuristic algorithms to real-World optimization problems. *Swarm and Evolutionary Computation*, 64, pp. 100888. doi: <https://doi.org/10.1016/j.swevo.2021.100888>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, pp. n71. doi: <https://doi.org/10.1136/bmj.n71>
- Popescu, D.-A. (2025). An Enhanced Genetic Algorithm for Optimized Educational Assessment Test Generation Through Population Variation. *Big Data and Cognitive Computing*, 9(4), pp. 98. doi: <https://doi.org/10.3390/bdcc9040098>
- Riese, E. & Bälter, O. (2022). A Qualitative Study of Experienced Course Coordinators' Perspectives on Assessment in Introductory Programming Courses for Non-CS Majors. *ACM Transactions on Computing Education (TOCE)*, 22(4), pp. 1–29. doi: <https://doi.org/10.1145/3517134>
- Shang, Y., Luo, X., Wang, L., Peng, H., Zhang, X., Ren, Y., et al. (2023). Reinforcement Learning Guided Multi-Objective Exam Paper Generation. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)* (pp. 829–837). SIAM. doi: <https://doi.org/10.1137/1.9781611977653.ch93>
- Spaccapanico Proietti, G., Matteucci, M. & Mignani, S. (2020). Automated Test Assembly for Large-Scale Standardized Assessments: Practical Issues and Possible Solutions. *Psych*, 2(4), pp. 315–337. doi: <https://doi.org/10.3390/psych2040024>
- Triantaphyllou, E. (2000). *Multi-criteria Decision Making Methods: A Comparative Study*. Springer. doi: <https://doi.org/10.1007/978-1-4757-3157-6>
- Van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. Springer. doi: <https://doi.org/10.1007/0-387-29054-0>
- Van der Linden, W. J. (2022). Review of the shadow-test approach to adaptive testing. *Behaviormetrika*, 49(2), pp. 169–190. doi: <https://doi.org/10.1007/s41237-021-00150-y>
- Veldkamp, B. P. & Paap, M. C. S. (2017). Robust Automated Test Assembly for Testlet-Based Tests: An Illustration with Analytical Reasoning Items. *Frontiers in Education*, 2, pp. 63. doi: <https://doi.org/10.3389/feduc.2017.00063>
- Wang, H., Zhuge, Q., Sha, E. H.-M., Xia, J. & Xu, R. (2024). Exploring Multiple-Objective Optimization for Efficient and Effective Test Paper Design with Dynamic Programming Guided Genetic Algorithm. *Mathematical Biosciences and Engineering*, 21(3), pp. 3668–3694. doi: <https://doi.org/10.3934/mbe.2024162>
- Wang, Y. & Wang, X. (2022). Test Paper Automatic Generating Method Based on Hybrid Genetic Algorithm. In S.-C. Chu, J. C.-W. Lin, J. Li, & J.-S. Pan (Eds.), *Genetic and Evolutionary Computing* (pp. 599–609). Springer Nature Singapore. doi: https://doi.org/10.1007/978-981-16-8430-2_54
- Wu, Z., He, T., Mao, C. & Huang, C. (2020). Exam paper generation based on performance prediction of student group. *Information Sciences*, 532, pp. 72–90. doi: <https://doi.org/10.1016/j.ins.2020.04.043>

- Xu, L., Jiang, Z., Cai, F., Ouyang, J., Liu, H. & Cai, T. (2024). Optimizing a national examination for medical undergraduates via modern automated test assembly approaches. *BMC Medical Education*, 24(1), pp. 919. doi: <https://doi.org/10.1186/s12909-024-05905-1>
- Yang, J. (2015). Design on Generating Test Paper Based on Simulated Annealing Algorithm. In *2nd International Conference on Civil, Materials and Environmental Sciences* (pp. 689–693). Atlantis Press. doi: <https://doi.org/10.2991/cmcs-15.2015.186>
- Zhang, J., Wong, C., Giacaman, N. & Luxton-Reilly, A. (2021). Automated Classification of Computing Education Questions using Bloom's Taxonomy. In *Proceedings of the 23rd Australasian Computing Education Conference* (pp. 58–65). Association for Computing Machinery. doi: <https://doi.org/10.1145/3441636.3442305>