Paper-Level Computerized Adaptive Testing for High-Stakes Examination via Multi-Objective Optimization

Mingjia Li
limj@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Junkai Tong
202103150317@zjut.edu.cn
East China Normal University
Shanghai, China

Yiyang Huang
10235102470@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Yifei Ding
10235102499@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Hong Qian*
hqian@cs.ecnu.edu.cn
East China Normal University
Shanghai, China

Aimin Zhou
amzhou@cs.ecnu.edu.cn
East China Normal University
Shanghai, China

Abstract

Computerized Adaptive Testing (CAT) is a testing technique that accurately infers students' proficiency levels using a relatively small number of questions. Most existing CAT systems operate on a question-level adaptive paradigm, which is suitable for practice scenarios. However, in computerized standardized high-stakes examinations such as the GRE and GMAT, this paradigm faces several challenges: (1) the lack of comparability in exam results, (2) high implementation costs due to the reliance on real-time interactions and the financial burden of maintaining CAT testing system, and (3) the difficulty in balancing multiple factors of diagnosis quality, attribute coverage, and question exposure. To address these challenges, we propose a Paper-level Computerized Adaptive Testing (PCAT) and its corresponding evaluation method. PCAT divides an exam into multiple testing stages, where examinees adaptively receive test papers of varying difficulty based on their performance in previous stages. The paper assembly problem in PCAT is solved using a population-based multi-objective optimization (MOO) approach. PCAT offers several advantages: First, the paper-level adaptive mechanism ensures that the questions faced by examinees depend solely on their performance in the earlier stages, maintaining adaptability while enhancing the comparability of results across different examinees. Second, PCAT replaces the selection strategy module in traditional CAT with an assembly module, allowing computationally intensive tasks such as cognitive diagnosis and paper assembly to be completed offline before the exam, eliminating the need for real-time interactions. Additionally, the population-based MOO approach generates a set of high-quality solutions in one run, meeting the demands of frequent administration of standardized high-stakes exams like the GRE and reducing the financial burden of maintaining a large-scale CAT system. Finally, MOO naturally models multiple factors as separate objectives, enabling a balanced consideration of these factors and allowing exam administrators to customize the exam based on specific needs. Extensive experiments on four real-world datasets show that PCAT outperforms state-of-the-art (SOTA) CAT methods in terms of diagnosis quality, attribute coverage, and question exposure, while maintaining the same number of questions answered by examinees. These results highlight PCAT's potential in high-stakes examination settings.

CCS Concepts

• **Applied computing** → Education; • **Computing methodologies** → Machine learning.

Keywords

Computerized adaptive testing, Paper assembly, Multi-objective optimization, High-stakes examination

ACM Reference Format:
Mingjia Li, Junkai Tong, Yiyang Huang, Yifei Ding, Hong Qian, and Aimin Zhou. 2025. Paper-Level Computerized Adaptive Testing for High-Stakes Examination via Multi-Objective Optimization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3737073

1 Introduction

The landscape of intelligent education is rapidly evolving, marked by innovations in personalized learning methodologies [19, 43], foundation model enhanced intelligent tutoring systems [28, 38], and automated discovery of study laws [20, 32]. In this context, Computerized Adaptive Testing (CAT) has emerged as a pivotal approach [33, 34], offering a significant advancement over traditional assessment with the paper-and-pencil testing method. By dynamically adjusting the difficulty and content of questions based on a student's ongoing responses, CAT aims to deliver personalized and efficient diagnostic services tailored to individual learners.

Unlike traditional paper-and-pencil tests, CAT interacts with examinees in a round-by-round manner, adaptively selecting questions tailored to the examinee's proficiency level to achieve rapid and accurate assessments of their abilities. As illustrated in Figure 1(b), a typical CAT system comprises several key components and follows a specific workflow: in each interaction round, a cognitive diagnosis model (CDM) [6, 9, 16, 17, 21, 22, 24, 29, 31, 36, 46],

*Hong Qian is the corresponding author.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737073

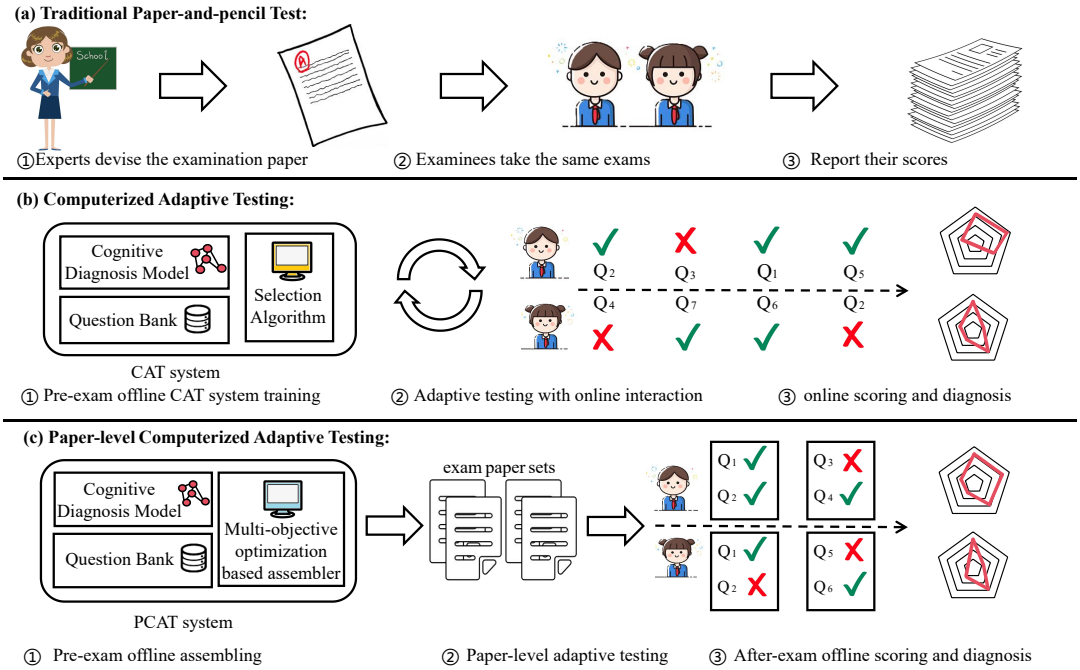

Figure 1: The comparison of paradigms for (a) Traditional paper and pencil test, (b) Computerized adaptive testing and (c) Paper-level computerized adaptive testing.

often referred to as the student profile module, estimates the examinee's proficiency level based on their current responses. Subsequently, the selection module leverages this proficiency estimate to choose the next most suitable question for the examinee.

Most existing CATs operate on a question-level adaptive paradigm, which has been proven to be effective in providing a customized and efficient testing experience for students during practice tests. However, for high-stakes examination scenarios, such as the GRE and GMAT, existing CAT approaches face several challenges:

- **Lack of Comparability**: Providing entirely different questions to different examinees hampers the ability to make horizontal comparisons between candidates. This reduces the transparency and persuasiveness of the exam results, which is particularly critical in high-stakes examination settings.
- **High Implementation Costs**: Existing CAT systems rely on real-time interactions with examinees, requiring frequent updates to the CDM. This not only prolongs the examination duration but also increases the financial burden of maintaining a large-scale CAT system, especially in environments with unstable network conditions or limited computational resources. Additionally, the need for frequent administration of standardized exams such as the GRE further exacerbates these costs, as traditional CAT systems struggle to efficiently generate and manage large volumes of test papers.
- **Difficulty in Balancing Multiple Factors**: Traditional CAT systems often prioritize test effectiveness and efficiency while paying less attention to other factors, such as attribute coverage and question exposure. Balancing these factors is essential for ensuring the quality, diversity and safety of high-stakes exams.

To address these challenges, this paper introduces Paper-level Computerized Adaptive Testing (PCAT), which is designed specifically for computerized high-stakes examinations. As depicted in Figure 1(c), PCAT divides the exam into multiple testing stages, where examinees adaptively receive papers of varying difficulty based on their performance in previous stages. Unlike traditional CAT, which utilizes a question-level adaptive paradigm, PCAT replaces the online selection strategy with an offline paper assembly module. This module leverages a population-based MOO approach to generate high-quality papers that balance multiple factors, such as diagnosis quality, attribute coverage (attribute refers to the knowledge concept associated with questions), and question exposure.

The PCAT approach offers several advantages over traditional CAT systems. First, by adopting a paper-level adaptive mechanism, PCAT ensures that the questions faced by the examinees depend solely on their performance in earlier stages. This maintains the adaptive nature of the test while enhancing the comparability of results across different examinees, thereby improving the transparency and persuasiveness of the exam outcomes. Second, PCAT shifts computationally intensive tasks, such as CDM update and question selection, to an offline phase of paper assembly before the exam. This eliminates the need for real-time interactions, which ensures the controllability of the examination process and enhances its robustness against poor conditions such as equipment and network issues. Furthermore, the population-based MOO approach generates multiple sets of high-quality test papers in a single run, meeting the demands of frequent exam administration. Together, these features significantly reduce the implementation costs of applying CAT in computerized high-stakes examinations. Finally, MOO naturally

models multiple factors as separate objectives, allowing a balanced consideration of diagnosis quality, attribute coverage, and question exposure. This flexibility allows exam administrators to customize paper assembly based on specific needs, such as prioritizing lower question exposure for exams with a high number of participants. Extensive experimental results on four real-world datasets demonstrate that the PCAT scheme, when compared with the SOTAs in existing CAT paradigm, achieves superior performance in terms of diagnosis quality, attribute coverage, and question exposure, while maintaining the same number of questions answered by the examinees. *Specifically, PCAT outperforms the SOTAs by an average of 5% in terms of accuracy across four datasets.* These findings underscore the potential of PCAT as a viable and effective solution to CAT in high-stakes examination settings.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work. Section 3 formalizes the task of PCAT and introduces a novel evaluation method tailored for PCAT. Section 4 presents the technical details of the PCAT framework, including solution representation, fitness evaluation, and the MOO workflow. Section 5 describes the experimental setup, datasets, and comparison methods, followed by a detailed analysis of the results. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2 Related Work

2.1 Computerized Adaptive Testing

Traditional CAT methodologies can be broadly categorized into statistical-based methods and machine learning-driven approaches. Statistical frameworks, such as Fisher Information (FI) [23] and Kullback-Leibler Information (KLI) [4], form the foundation of classical CAT systems. FI selects questions to maximize the local information around the examinee's current proficiency estimate, while KLI enhances robustness in early testing stages by integrating global divergence metrics. Recent extensions like Multi-Objective Optimization of Item Selection (MOOIT) [25] leverage genetic algorithms to model the trade-off between test length and precision, explicitly balancing estimation accuracy and testing efficiency through Pareto-optimal solutions. However, these methods remain constrained by rigid psychometric assumptions and struggle to address multifaceted challenges like attribute coverage and question exposure control.

In recent years, machine learning-driven approaches have revolutionized CAT by automating adaptive policies through data-driven learning and optimization. Reinforcement learning (RL)-based methods, such as graph-enhanced multi-objective CAT (GMOCAT) [37] and neural CAT (NCAT) [48], exemplify this shift: GMOCAT integrates graph neural networks with multi-objective RL to optimize multiple factors simultaneously, while NCAT employs transformer architectures to capture complex examinee-question interactions under an RL formulation of CAT. Meta-learning techniques, represented by bilevel optimization-based CAT (BOBCAT) [10], adopt bi-level optimization to train model-agnostic selection policies, allowing rapid adaptation across diverse testing scenarios. Search and optimization-oriented methods further enhance scalability and efficiency: bounded estimation CAT (BECAT) [49] redefines CAT as a gradient-driven subset selection problem with theoretical error guarantees; search-efficient CAT (SECAT) [12] reduces time complexity via hierarchical indexing and multi-round search strategies; and unified adaptive testing (UAT) [45] introduces a hierarchical structure search framework to unify diverse testing formats under theoretical guarantees. Complementing these, active learning-inspired methods like model-agnostic adaptive testing (MAAT) [1] prioritize question diversity through uncertainty sampling and representativeness metrics.

Despite their advancements, the existing methods adhere to the *one-for-each* paradigm, generating unique question sequences in real time for individual examinees [33]. Although effective for personalized assessments, this approach introduces critical limitations in high-stakes standardized exams: prohibitive implementation costs from real-time computation and comparability concerns.

2.2 Population-based Multi-Objective Optimization

MOO addresses problems with conflicting objectives by seeking Pareto-optimal solutions that balance trade-offs among them. Classical MOO methods include genetic algorithms such as NSGA-II [7], which employs non-dominated sorting and crowding distance to maintain solution diversity, and decomposition-based approaches like MOEA/D [47] that transform MOO into scalar subproblems. These population-based methods excel at generating diverse solution sets in a single optimization run, making them suitable for scenarios requiring multiple high-quality alternatives.

In the context of CAT, several studies have explored the application of MOO to balance multiple factors such as diagnosis accuracy, test efficiency, and question diversity. For instance, MOO of Item Selection (MOOIT) [25] employs genetic algorithms to model the trade-offs between test length and precision, explicitly balancing estimation accuracy and testing efficiency through Pareto-optimal solutions. Similarly, reinforcement learning-based methods like GMOCAT [37] integrate MOO with adaptive policies to optimize multiple factors simultaneously. These approaches, while effective in balancing multiple objectives, are still confined to the traditional CAT paradigm, where the optimization process is tailored to individual examinees in real-time.

In addition, in the realm of test assembly, which also involves the construction of test papers, existing approaches predominantly focus on single-stage scenarios without adaptive mechanisms. Notably, parallel test assembly methods [3, 14, 18] generate multiple papers simultaneously, sharing similarities with our setting in terms of batch processing. However, these approaches remain limited to non-adaptive test construction. Specifically, AR-DMOEA [42], proposes a MOO-based approach for on-the-fly assembled multistage adaptive testing. Although AR-DMOEA involves multiple testing stages, it still adheres to the *one-for-each* paradigm, where different examinees receive distinct question sequences, thereby still facing the challenge of ensuring comparability across students. Prior studies have investigated examination generation systems grounded in performance mechanisms [41]. A notable advancement in this domain is MOEPG [30], which employs a reinforcement learning-driven paradigm to balance diverse test-design objectives. Nevertheless, these approaches predominantly focus on student-level

adaptation, while systematically overlooking the critical need for multi-phase pedagogical adaptation.

In summary, while MOO has shown promise in addressing the multifaceted challenges of CAT, the current applications are primarily confined to the question-level adaptive paradigm. The challenge remains to develop methods that can balance multiple objectives while ensuring the comparability of high-stakes examinations.

3 PCAT Formalization and Evaluation

3.1 Task Formalization

Given the question bank $\boldsymbol{Q} = \{Q_1, Q_2, \ldots, Q_n\}$, in which $C = \{C_1, C_2, \ldots, C_d\}$ denotes the attribute set and a $Q\text{-}matrix \in \{0,1\}^{n \times d}$ denotes the attributes assessed by each question. The task of a PCAT system is to generate several sets of test papers based on the response logs $\mathcal{R} = \{R_1, R_2, \ldots, R_\ell\}$ collected during the question evaluation phase for $\boldsymbol{Q}$. The generated sets of papers are denoted as $\mathcal{P} = \{P_1, P_2, \ldots, P_m\}$, Here, each set of papers $P_i \in \mathcal{P}$ consists of k testing stages, which collectively constitutes the complete process of a k-stage examination. In a k-stage examination, the first testing stage contains a single test paper, while each subsequent testing stage comprises multiple test papers of varying difficulty levels. In an examination, the test paper presented to a student at each stage is adaptively determined based on their performance in the preceding stages, following a rule-based or heuristic strategy. Formally, for a set of papers P_i, the t-th testing stage ($t \in \{2, \ldots, k\}$) can be represented as $P_{i,t} = \{P_{i,t,1}, P_{i,t,2}, \ldots, P_{i,t,u}\}$, where each $P_{i,t,j} \subseteq \boldsymbol{Q}$ refers to a paper. The total number of questions involved in P_i is denoted as $|P_i|$. The set of response logs $\mathcal{R}$ contains ℓ interaction records, where each record is represented as a triple (S_i, Q_j, r_{ij}). Here, $S_i \in \mathcal{S}$ denotes the i-th student in the student set $\mathcal{S}$, Q_j represents the j-th question from the question bank $\boldsymbol{Q}$, and $r_{ij} \in \{0, 1\}$ indicates the correctness of the student's response to the question, with $r_{ij} = 1$ if the response is correct and $r_{ij} = 0$ otherwise.

3.2 PCAT Evaluation

Traditional evaluation methods for CAT are not directly applicable to PCAT. In conventional CAT, given a sparse set of response logs $\mathcal{R}$, the evaluation process typically involves partitioning the students into training and test sets. For each test student, the questions in $\mathcal{R}$ with recorded responses are further divided into a support set and a query set. After training the CDM and the selection strategy in the training set, the CAT system selects a subset of questions from the support set for the test student, collects their responses, and infers their proficiency level using the CDM. The performance is then evaluated based on the prediction accuracy on the query set. However, this approach cannot be directly applied to PCAT because, in PCAT, the test papers are pre-assembled, and it is possible that some questions in the test papers may not have corresponding response records in $\mathcal{R}$ for certain test students. This issue renders the traditional evaluation method infeasible for PCAT.

To address this challenge, we propose a novel evaluation approach tailored to PCAT. Specifically, we first employ a heuristic algorithm (detailed in Section 5.1) to extract a dense subset $\mathcal{R}_D$ from the sparse response logs $\mathcal{R}$. In $\mathcal{R}_D$, every student has a recorded response for every question, thereby eliminating the issue of missing response records during evaluation. The complete evaluation process is as follows: Given the dense response logs $\mathcal{R}_D$, the students are randomly partitioned into training and test sets. The paper assembly process is performed on the training set to generate the pre-assembled test papers. For each test student, the examination is conducted following the aforementioned multi-stage paper-level adaptive testing paradigm, where the test papers are adaptively selected based on the student's performance in previous stages. The CDM is then applied to infer the student's proficiency level based on their examination results. Finally, the evaluation metrics are computed based on the prediction performance on the query set (detailed in section 5.3), ensuring a fair and comprehensive assessment of the PCAT system's performance. This approach not only overcomes the limitations of traditional CAT evaluation but also aligns with the unique characteristics of PCAT, providing a robust framework for evaluating its effectiveness in high-stakes examination settings.

4 Methodology: Paper-Level Computerized Adaptive Testing (PCAT)

In this section, we present the technical details of PCAT. The general framework of MOO consists of several key components, namely, solution representation, fitness evaluation, initialization, mutation and crossover, selection mechanism and termination. The core innovation of PCAT lies in the representation of solutions, where we introduce a dual-layer encoding for questions that combines the question type and difficulty level. The fitness evaluation is conducted through a simulated paper-level adaptive testing process, assessing diagnostic quality, attribute coverage, and question exposure. The following subsections detail the solution representation, fitness evaluation, and the complete workflow of PCAT.

4.1 Solution Representation

The representation of individual solutions is a fundamental aspect of evolutionary algorithms, as it directly impacts the efficiency and interpretability of the optimization process. Existing approaches that apply MOO to CAT and test assembly [25, 37, 42] typically use binary representations, where a solution is encoded as a long 0-1 vector, with each index indicating the selection of each question in the question bank $\boldsymbol{Q}$. Although simple, this representation lacks educational semantics and becomes inefficient as the question pool grows, leading to significant memory wastage and computational overhead due to the sparse nature of the vector.

To address these limitations, we propose a novel dual-layer representation of questions. In this approach, each question is encoded as a *type-difficulty pair*: the first dimension represents the question type, which is an integer $\tau \in \{1, 2, \ldots, \eta\}$ indicating the combination of attributes being assessed (where η is the total number of question types, and each type corresponds to a unique combination of attributes from the attribute set C), and the second dimension represents the difficulty level, which is a real-valued coefficient $\delta \in \mathbb{R}$. Formally, a question is represented as a tuple (τ, δ). During the evolutionary process, a solution, i.e., a set of papers can be represented as a matrix $\boldsymbol{M_Q}$ with a size of $|P_i| \times 2$. The decoding of $\boldsymbol{M_Q}$ involves mapping each row $(\tau_i, \delta_i), i = 1, 2, \ldots, h$ in $\boldsymbol{M_Q}$ to a specific question Q_i from $\boldsymbol{Q}$ by a neighborhood projection process,

i.e., selecting the question of type τ_i whose difficulty level is closest to δ_i. Formally, this can be expressed as:

$$Q_i = \arg\min_{q \in \mathcal{Q}_{\tau_i}} |\delta^q - \delta_i|, \tag{1}$$

where $\mathcal{Q}_{\tau_i} \subseteq \mathcal{Q}$ is the subset of questions in $\mathcal{Q}$ with type τ_i, and δ^q denotes the difficulty level of question q.

The proposed dual-layer representation offers several advantages. First, it is highly compact, significantly reducing the solution space compared to traditional binary representations. For example, a test paper with h questions requires only $2h$ values to represent, as opposed to n values in the binary representation (where $n \gg h$). Second, it incorporates educational semantics, making solutions more interpretable and aligned with the goals of educational assessment. Third, it improves computational efficiency by avoiding the memory and processing overhead associated with sparse binary vectors. These benefits make the dual-layer representation particularly suitable for large-scale, high-stakes testing scenarios, where both efficiency and interpretability are critical.

4.2 Fitness Evaluation

The fitness evaluation process simulates the multi-stage paper-level adaptive mechanism to evaluate the fitness of a solution in multiple factors. Specifically, students begin with the same test paper in the first stage. For subsequent stages, the system adaptively assigns test papers of varying difficulty levels based on the students' performance in previous stages. Specifically, students are ranked according to their diagnosed mastery levels and divided into equal-sized groups, with each group assigned a test paper of a corresponding difficulty. After all stages are completed, the response logs from the examination are diagnosed to produce the final metrics to evaluate the overall performance of the paper set.

In the optimization process of this paper, we adopt a two-stage PCAT scenario, inspired by the design of the GRE examination. Specifically, given a set of papers P_i with $|P_i|$ questions, it will be divided into 3 papers with different difficulty levels (easy, medium, hard), each with $|P_i|/3$ questions. All the students in the training set are presented with the medium paper in the first stage. In the second stage, based on their performance in the previous stage, the top performing half of the students are assigned the hard test paper, while the remaining half receive the easy one. Ultimately, each student completes a total of $Step = 2/3|P_i|$ questions.

The evaluation metrics focus on three key aspects: diagnostic quality, attribute coverage, and question exposure. Specifically, we use the area under the curve (AUC) and accuracy (ACC) on the training set of a CDM to measure the diagnostic quality of the solution P_i. These metrics are computed as follows:

$$\text{AUC} = \int_0^1 \text{TPR}(f) \cdot \text{FPR}(f)\, \mathrm{d}f, \ \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2}$$

where TPR and FPR denote the true positive rate and false positive rate, respectively, and TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

For attribute coverage (COV), we calculate the average number of attributes covered by the questions selected for each student. This is defined as:

$$\text{COV} = \frac{1}{|\mathcal{S}_\mathcal{T}|} \sum_{s \in \mathcal{S}_\mathcal{T}} \left| \frac{1}{|C|} \bigcup_{q \in \mathcal{Q}_s} \mathcal{Q}\text{-}matrix(q) \right|, \tag{3}$$

where $\mathcal{S}_\mathcal{T}$ is the set of students in the training set, $\mathcal{Q}_s$ is the set of questions assigned to student s, and $\mathcal{Q}\text{-}matrix(Q_j)$ denotes the attributes assessed by question Q_j.

For question exposure (EXP), we measure the frequency with which each question appears across all students' test papers, as:

$$\text{EXP} = \frac{1}{|\mathcal{S}_\mathcal{T}|} \sum_{s \in \mathcal{S}_\mathcal{T}} \left(\frac{\sum_{q \in \mathcal{Q}_s} (N_q / |\mathcal{S}| - r)^2}{r} \right), \tag{4}$$

where $r = Step/n$, N_q denotes the number of responses for question q in the original response logs $\mathcal{R}$. $Step$ denotes the testing steps, that is, how many questions a student will answer in a PCAT exam.

These metrics collectively provide a comprehensive evaluation of the test paper set's effectiveness in terms of diagnostic quality, attribute diversity, and question bank safety.

4.3 Multi-Objective Optimization in PCAT

Our approach employs an MOO approach based on NSGA-II [7], utilizing type-difficulty pair vectors to represent individual solutions. This framework ensures efficient exploration of the solution space while adhering to the objectives and constraints of the target.

4.3.1 Solution Initialization. Each individual solution is represented as a vector of type-difficulty pairs, where each pair corresponds to a specific question. The type is an integer $\tau \in \{1, 2, \ldots, \eta\}$ representing the combination of attributes being assessed, and the difficulty is a real-valued coefficient $\delta \in [\delta_{\min}, \delta_{\max}]$. During initialization, the type for each pair is randomly selected from the set of question types, ensuring that there are no duplicate types within a single solution. The difficulty is initialized uniformly within the predefined range $[\delta_{\min}, \delta_{\max}]$, which is derived from CDM. This ensures diversity and feasibility in the initial population.

4.3.2 Crossover and Mutation. Crossover and mutation operations are designed to explore the solution space effectively. For crossover, we use the traveling salesman crossover operator [11] and the 2-point crossover operator [39] for type and difficulty, respectively. For mutation, a two-point mutation strategy [40] is applied, where both the type and difficulty of a pair are perturbed with a predefined probability. The type mutation is limited to valid question types, while the difficulty mutation remains within the range $[\delta_{\min}, \delta_{\max}]$. Formally, for a pair (τ, δ), the mutation process can express as

$$\tau' = \tau + \Delta_\tau, \quad \delta' = \delta + \Delta_\delta, \tag{5}$$

where Δ_τ and Δ_δ are random perturbations.

4.3.3 Selection Mechanism. The selection process combines the non-dominated sorting and crowding distance to maintain a balance between convergence and diversity. Non-dominated sorting assigns a rank to each solution based on its dominance relationship, where a solution P_i dominates P_j if P_i is superior in at least one objective and not worse in any other. Crowding distance measures the density of solutions in the objective space, prioritizing solutions in less

Algorithm 1 Multi-Objective Optimization Framework for PCAT

Input: Student set $\mathcal{S}$, Question bank $\mathcal{Q}$, the max iteration T
Initialization:
Randomly generate a population of solutions $\mathcal{P}$, where a solution P_i denotes a set of papers. Iteration indicator $itr \leftarrow 0$.
while $itr < T$ **do**
  Randomly sample parents from $\mathcal{P}$, conduct crossover and mutation to generate new population $\mathcal{P}^*$
  for P_i **in**$\mathcal{P}$ **do**
    for q **in**P_i **do**
      Decode the type-difficulty pair of q into a question $Q_j \in \mathcal{Q}$ via neighborhood projection
    end for
  end for
  Transform P_i into three papers (*easy*, *medium* and *hard*).
  for s **in** $\mathcal{S}_\mathcal{T}$ **do**
    $rank \leftarrow$ the percentage ranking of student s' mastery level on the *medium* paper
    if $rank > 0.5$ **then**
      Adaptive to the *hard* paper
    else
      Adaptive to the *easy* paper
    end if
  end for
  Calculate four objectives AUC, ACC, COV and EXP according to the responses.
  Select new population $\mathcal{P}$ from $\mathcal{P}^*$ according to the fitness on four objectives and the crowding distance.
  $itr = itr + 1$
end while
Output: The final population $\mathcal{P}$.

crowded regions. For a solution P_i, the crowding distance $d_j(P_i)$ for the j-th objective is computed as:

$$d_j(P_i) = \frac{f_j(P_{i+1}) - f_j(P_{i-1})}{f_j^{\max} - f_j^{\min}}, \tag{6}$$

where P_{i+1} and P_{i-1} are adjacent solutions, and $f_i^{\max}$ and $f_i^{\min}$ are the maximum and minimum values of the i-th objective. The total crowding distance for a solution P_i is computed as: $D(P_i) = \sum_{j=1}^{o} d_j(P_i)$, where o is the number of objectives. Solutions with higher ranks and larger crowding distances are preferential for the next generation, ensuring a diverse and high-quality population. The complete workflow of PCAT refers to Algorithm 1.

5 Experiments

In this section, we conduct extensive experiments on four real-world datasets to evaluate the performance of PCAT. The code is available at https://github.com/ECNU-ILOG/PCAT.

5.1 Data Processing and Experimental Setup

We utilize four real-world educational datasets in our experiment, namely MOOCRadar (**MOOC**) [44], Assistment 2017 (**ASSIST**) [5], ProbabilityAndStatistic (**PROB-STA**) [13] and ComputationalThinking (**COM-TH**) [13]. These four datasets are sourced from different online education platforms and have been widely used in the fields of cognitive diagnosis, knowledge tracing and CAT.

Table 1: Statistics of the densified datasets

Datasets	ASSIST	MOOC	PROB-STA	COM-TH
# Students	120	1,230	105	2,433
# Questions	83	181	263	272
# Attributes	43	180	63	57
# Records	9.9K	220K	28K	1.6M
Attribute Per Question	1.51	1.64	1.00	1.01
Positive Label Rate	0.68	0.95	0.37	0.77

5.1.1 Datasets Densification. As disclosed in Section 3.2, the original datasets are sparse, with each student responding to only a small subset of questions, making them incapable of directly evaluating the performance of PCAT. To address this, our aim is to transform the sparse response matrix into a dense one while maximizing its size. This problem is essentially a variant of the densest subgraph problem, which is known to be NP-hard. To efficiently densify the datasets, we propose a heuristic algorithm that iteratively removes students or questions with the fewest responses until the resulting submatrix is dense. Specifically, at each step, we remove either the student with the fewest responses or the question with the fewest responses, whichever contributes more to the sparsity. This process continues until the remaining submatrix is dense, ensuring that every student in the subset has answered every question in the subset. The detailed procedure is outlined in Algorithm 2 in Appendix A. The statistics of the densified datasets are summarized in Table 1.

5.1.2 Experimental Setup. In our experiment, students are partitioned into training and test sets with a 80%-20% split. The training data are used to train the CDM and generate the sets of testing papers in PCAT, while in CAT, they are used to train the CDM and the selection strategy. In the testing phase, students in the test set are treated as new examinees for the CAT system. Both the baseline CAT methods and our proposed PCAT algorithm are agnostic to the students' response records. The response labels for the selected questions are only revealed after the selection is finalized. We set the test length $Step$ (i.e., the number of questions answered by each student) the same for PCAT and the CAT baselines, allowing us to compare the performance of CAT and PCAT in a relatively fair way. Specifically, we evaluate the algorithms with $Step = 10$ and $Step = 20$, and experiments in each setting are repeated 5 times.

5.2 Implementation of PCAT and Compared Methods

In this paper, we adopt a 2-stage PCAT setting inspired by the GRE examination model, as described in Section 4.2. Key hyperparameters and configurations are as follows:

- MOO: In the process of MOO, we adopt the maximum evolutionary iteration $T = 100$ and population size = 80.
- Cognitive Dianosis Model: We adopt the neural cognitive diagnosis model [35] as the student profile module.
- Learning Rate: The learning rate for CDM training is set to 0.05.
- Batch Size: The batch size for CDM training is set to 512.

- Network Architecture: The neural network in CDM consists of two hidden layers with 128 and 64 units, respectively.
- Optimizer: We use the Adam optimizer [15] in CDM.

We choose the following methods in the field of CAT as comparison algorithms. For detailed information about these methods, please refer to the **Related Work** Section 2.1.

Statistical-based CAT.

- **Random:** Randomly select questions from Q for students. It is worth noting that **Random** can also be regarded as the most naive paper assembly method.
- **MFI** [23]: Select the question that provides the maximum Fisher information for students' mastery level correction.
- **KLI** [4]: Select the question which brings the maximum Kullback-Leibler information to the difference of students' mastery levels.

Machine learning-driven CAT.

- **MAAT** [1]: Select questions that balance the gain of the students' mastery levels and the coverage of attributes.
- **BOBCAT** [10]: Minimize the cross-entropy loss in the inner layer to estimate the students' mastery level, and optimize the selection algorithm and global parameters in the outer layer.
- **NCAT** [48]: Employ attention neural networks to model the interaction function, and then use Q-learning to select questions.
- **BECAT** [49]: Adopt the similarity between pairs of questions and utilize a greedy way to find the optimal subset of questions.
- **GMOCAT** [37]: Use graph neural networks to learn relation-aware embeddings to aggregate domain information. An reinforcement learning-based selection policy is introduced to optimize three objectives: quality, diversity and novelty.

All experiments are conducted on a Linux server with two 3.00GHz Intel Xeon Gold 6354 CPUs and one RTX3090 GPU. The code is written in Pytorch [27] and MOO is implemented using Pymoo [2].

5.3 Evaluation Metrics

As described in Section 4.2, the evaluation of our PCAT framework focuses on three key aspects: diagnostic quality, attribute coverage, and question exposure, encompassing four objectives: AUC, ACC, COV and EXP. Specifically, the definitions of AUC and ACC are the same as Eq. (2). The metric of COV can be expressed as follows:

$$\text{COV} = \frac{1}{|\mathcal{S}_\mathcal{E}|} \sum_{s \in \mathcal{S}_\mathcal{E}} \left| \frac{1}{|C|} \bigcup_{q \in Q_s} Q\text{-}matrix(q) \right| . \tag{7}$$

The metric of EXP is defined as:

$$\text{EXP} = \frac{1}{|\mathcal{S}_\mathcal{E}|} \sum_{s \in \mathcal{S}_\mathcal{E}} \left(\frac{\sum_{q \in Q_s} (N_q/|\mathcal{S}| - r)^2}{r} \right), \tag{8}$$

where $\mathcal{S}_\mathcal{E}$ denotes the set of students in the test set, and other notations are the same as in Eq. (3) and Eq. (4) respectively.

Furthermore, to analyze the implementation cost associated with the execution time of each algorithm, we categorize the total wall-clock execution time into offline and online components based on the requirement for real-time student interaction. We emphasize the online execution time because, in high-stakes examinations, it is crucial to consider its duration to ensure that the exam process remains manageable and controlled.

5.4 Experimental Results

5.4.1 Diagnostic Quality. We conduct student performance prediction experiments under all baseline methods and the PCAT framework to compare the quality results. Table 2 presents the specific information of the quality metrics, i.e., AUC and ACC, for each algorithm, with results shown for $Step = 10$ and $Step = 20$. It is worth noting that in PCAT, the final output is a population of the set of papers. From this population, we select the solution with the highest $AUC+ACC$ on the training set and report its AUC and ACC on the test set. All results are independently repeated five times, and we report the mean ± standard deviation.

From Table 2, it can be observed that PCAT outperforms all baseline methods in terms of quality objectives. This result is reasonable because, compared to the single-step approach, the global search framework of PCAT considers the global effect, which is more likely to achieve better expectations for students' ability estimation. On most datasets, the PCAT framework consistently achieves optimal or near-optimal results in both AUC and ACC. This indicates that the multi-objective framework effectively balances the validity of each objective, providing a robust and accurate assessment of students' proficiency levels.

Specifically, for the **ASSIST** dataset, PCAT achieves the highest AUC values of 84.59% and 87.19% for $Step = 10$ and $Step = 20$ respectively, which are significantly higher than the second-best results of 81.63% and 83.01%. Similarly, for ACC, PCAT achieves 77.52% and 79.97% for $Step = 10$ and $Step = 20$, respectively, outperforming all other methods. On the **MOOC** dataset, PCAT achieves competitive results, with AUC values of 59.01% and 61.51% for $Step = 10$ and $Step = 20$ respectively, and ACC values of 77.19% and 86.86%. These results are comparable to or better than the best-performing baselines, such as MFI and GMOCAT. For the **PROB-STA** dataset, PCAT achieves the highest AUC value of 61.97% for $Step = 20$, slightly outperforming the second-best result from MFI (61.88%). In terms of ACC, PCAT achieves 62.37% for $Step = 20$, which is close to the best result from MFI (62.82%). Finally, on the **COM-TH** dataset, PCAT achieves the highest AUC value of 61.00% for $Step = 20$, outperforming all other methods. For ACC, PCAT achieves 85.26%, which is comparable to the best result from KLI (85.63%). *It is worth noting that PCAT outperforms the SOTAs by an average of 5% in terms of ACC across four datasets.*

Overall, the experimental results suggest that the PCAT framework consistently achieves superior or competitive performance across different datasets and $Step$ sizes, highlighting its effectiveness in accurately assessing students' proficiency levels.

5.4.2 Attribute Coverage and Question Exposure. The diversity of questions chosen for students is an important indicator, reflecting the range of attributes covered by the questions and ensuring a comprehensive assessment. It is worth noting that the COV and EXP metrics on different datasets are not on the same scale. Therefore for ease of visualization, we present the ratios of each method's performance relative to PCAT on the respective dataset, denoted as EXP Ratio and COV Ratio. Figure 4 compares the attribute coverage results at $Step = 20$ for the baselines and PCAT. PCAT performs well across all four datasets, consistently outperforming all baseline methods. This superior performance is attributed to PCAT's intrinsic objective balancing and its performance-guaranteed search

Table 2: The AUC and ACC comparison on four real-world datasets. The best performance is in bold, while the second best value is underlined. "-" indicates the method cannot give results within 7 hours measured by online time. "*" indicates statistically significant improvement compared with the second best (measured by t-test) with p-value < 0.05.

Dataset	ASSIST		MOOC		PROB-STA		COM-TH	
Metric	AUC		AUC		AUC		AUC	
Step	10	20	10	20	10	20	10	20
Random	79.15±0.45	78.07±1.17	59.47±0.52	61.10±0.33	60.17±2.23	60.88±2.08	57.61±0.41	59.07±0.60
MFI	78.44±0.90	74.78±1.75	**61.30±0.45**	60.51±1.24	**61.76±1.40**	61.88±1.25	59.24±0.29	59.19±0.86
KIL	78.90±0.56	77.40±2.43	58.65±0.75	59.66±0.66	61.01±1.45	61.17±1.81	56.47±0.36	56.46±0.54
MAAT	80.11±1.02	79.64±1.67	-	-	60.38±1.00	61.11±1.02	-	-
NCAT	78.43±1.17	78.20±2.07	-	-	61.09±1.53	61.19±1.95	-	-
BOBCAT	78.41±0.90	75.75±2.29	59.99±0.43	60.28±0.87	61.46±1.01	61.09±1.71	56.66±0.69	56.37±0.96
BECAT	81.43±0.73	84.03±1.10	-	-	-	-	-	-
GMOCAT	81.63±1.06	83.01±1.25	58.84±0.50	59.69±0.56	60.60±1.20	60.62±1.32	**59.60±0.93**	58.47±0.33
PCAT	**84.59±0.79***	**87.19±0.76***	59.01±0.79	**61.51±0.79**	60.10±1.10	**61.97±1.11**	58.26±0.49	**61.00±0.60**
Metric	ACC		ACC		ACC		ACC	
Step	10	20	10	20	10	20	10	20
Random	69.00±0.45	69.02±0.99	66.12±3.29	68.84±3.18	60.9±2.03	61.72±1.66	85.01±0.68	85.20±0.61
MFI	69.02±1.15	65.86±2.55	68.43±3.59	74.19±4.07	**62.68±0.61**	**62.82±0.35**	84.89±0.64	85.22±0.57
KIL	68.30±1.53	68.00±3.49	65.23±3.65	67.63±3.40	62.01±0.71	62.04±0.94	85.16±0.66	**85.63±0.56**
MAAT	70.83±0.68	71.64±0.90	-	-	61.10±0.57	62.56±1.71	-	-
NCAT	69.55±0.89	69.44±3.05	-	-	61.63±0.95	61.88±0.84	-	-
BOBCAT	68.42±0.53	65.36±1.79	68.46±3.25	69.87±3.69	62.22±1.00	62.01±1.98	84.83±0.85	84.92±0.92
BECAT	71.51±0.84	73.94±0.70	-	-	-	-	-	-
GMOCAT	73.48±1.33	75.45±0.96	64.39±3.46	79.70±2.07	58.37±1.48	57.18±4.28	85.12±0.59	85.13±0.53
PCAT	**77.52±0.62***	**79.97±1.05***	**77.19±4.57***	**86.86±1.22***	56.15±2.99	62.37±1.51	**85.21±0.50**	85.26±0.55

algorithm. In contrast, baseline methods such as KLI, MFI, and even the multi-objective GMOCAT, which includes coverage as an objective, perform slightly worse than PCAT. We believe this is because algorithms like KLI, MFI, and GMOCAT focus on local effects through iterative selection, whereas PCAT takes a global approach, considering overall performance. While gradual selection may offer more personalized assessments, PCAT provides stronger fairness (e.g., testing all students on a diverse range of questions) and efficiency (e.g., no stepwise selection or calculation).

In addition to quality and coverage metrics, which represent assessment accuracy and question diversity, respectively, we also consider the exposure metric to ensure question bank security by limiting the number of times a question is exposed. Figure 5 shows the question exposure results at $Step = 20$ for each baseline method and PCAT. Random selection, though offering a wide range of question selection, violates the intent of CAT and leads to some loss in quality and coverage. Despite its diversity, it sets the upper benchmark for exposure. MFI and KLI, which focus solely on accuracy, show minimal improvements in coverage and exposure. Multi-objective methods like GMOCAT improve all three objectives simultaneously but lack a holistic approach due to their iterative nature. In comparison, PCAT achieves the lowest exposure, proving the effectiveness of its multi-objective optimization algorithm and demonstrating that directly optimizing exposure leads to reduced exposure overall. On the **MOOC** dataset, PCAT and compared methods exhibit equal EXP metrics. This observation is attributed to the unique structure of the MOOC dataset, wherein inherently uniform question exposure rates diminish the discriminative power of the EXP metric. Nevertheless, this metric is retained in the reported results for consistency in presentation.

The results highlight that PCAT not only maintains high diagnostic quality, but also ensures better attribute coverage and question exposure, making it a robust solution for high-stakes examinations.

5.4.3 Execution Time. The online execution times of each baseline and PCAT on the four datasets are shown in Table 3. The results clearly show the superiority of PCAT in terms of implementation cost. While others, such as MAAT and BECAT, exhibit substantial online execution times ranging from several minutes to over 7 hours, PCAT consistently completes tasks in near-zero time across all datasets. This significant time efficiency highlights the potential of PCAT in large-scale standardized examination settings.

5.4.4 Objective Comparison. During the training phase, a set of optimal non-dominated solutions is obtained. Since the PCAT framework involves four objectives, we divide them into groups for easier presentation of the resulting Pareto front: (1) AUC-EXP-COV and (2) ACC-EXP-COV. The results on **ASSIST** and **COM-TH** are shown in Figure 2 and 3, and the results of the other datasets are provided in Appendix B. In Figure 2 and 3, we plot EXP and COV on the x and y axes, respectively. Each individual solution is represented by a red dot, highlighting its position in the solution space. Additionally,

Table 3: The online time comparison when $Step = 20$.

Dataset	ASSIST	MOOC	PROB-STA	COM-TH
Random	1s	1s	1s	1s
MFI	18s	3min	25s	6min
KIL	5s	1min	5s	10min
MAAT	20min	>7h	50min	>7h
NCAT	<5s	>7h	3min	>7h
BOBCAT	3s	40s	3s	1min
BECAT	2h	>7h	>7h	>7h
GMOCAT	10min	2h	10min	4h
PCAT	1s	3s	1s	5s

a Pareto front surface is fitted to the data, with darker colors indicating lower AUC or ACC values and lighter colors representing higher values. As shown in Figure 2 and 3, the Pareto front forms an approximately smooth surface, which indicates that the PCAT framework effectively balances multiple competing objectives, providing a diverse set of high-quality solutions that can be tailored to specific examination needs.

These results demonstrate that the PCAT framework is capable of generating solutions that optimize multiple objectives simultaneously, ensuring a balanced consideration of question exposure, attribute coverage, and diagnostic quality. This flexibility allows exam administrators to customize the examination process based on their specific requirements, further enhancing the applicability of PCAT in high-stakes testing scenarios.

5.4.5 Hyper-parameters Analysis. To investigate the impact of hyper-parameters of MOO on the results of PCAT, we compare the AUC metric under different population size and the maximum number of iterations. Specifically, to study the effect of population size, we fix the maximum iteration and vary the population size. Similarly, to study the effect of maximum iteration, we fix population size and vary maximum iteration. The detailed results refer to Appendix C.

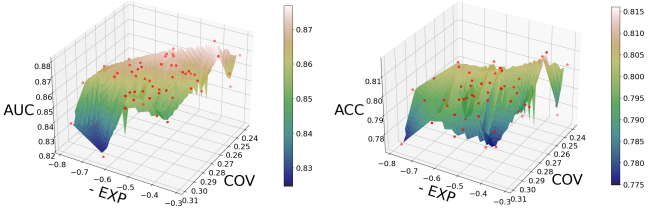

Figure 2: The Pareto front of the population assembled by PCAT on ASSIST when $Step = 20$.

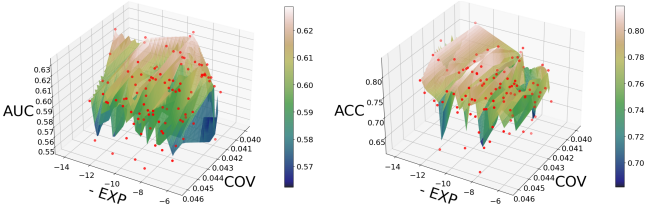

Figure 3: The Pareto front of the population assembled by PCAT on COM-TH when $Step = 20$.

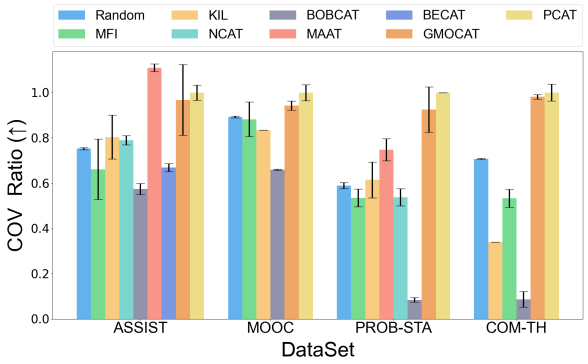

Figure 4: The attribute coverage comparison when $Step = 20$.

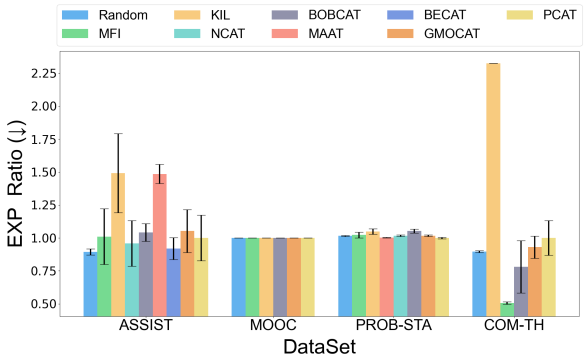

Figure 5: The question exposure comparison when $Step = 20$.

6 Conclusion

This paper introduces PCAT, a novel framework that addresses key limitations of traditional question-level adaptive testing in high-stakes exams. By adopting a paper-level adaptive mechanism and leveraging population-based MOO, PCAT enhances result comparability, reduces implementation costs, and balances multiple testing factors effectively. Experimental results demonstrate its superiority over SOTA methods in diagnosis quality, attribute coverage, and question exposure. Two promising directions for future research include: (1) developing dynamic, population-dependent metrics, such as question exposure rates adjusted based on the distribution of selected solutions; and (2) extending PCAT to more complex multi-stage testing scenarios. These advances could further improve the flexibility and applicability of PCAT.

Acknowledgments

We would like to express our sincere thanks to the anonymous reviewers for their constructive comments. The algorithms and datasets in the paper do not involve any ethical issue. This work is supported by the National Natural Science Foundation of China (No. 62476091).

References

[1] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. 2020. Quality meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing. In *Proceeding of the 20th IEEE International Conference on Data Mining*. Sorrento, Italy, 42–51.

[2] Julian Blank and Kalyanmoy Deb. 2020. Pymoo: Multi-Objective Optimization in Python. *IEEE Access* 8 (2020), 89497–89509.

[3] Xi Cao, Ying Lin, Dong Liu, Henry Been-Lirn Duh, and Jun Zhang. 2024. Large-Scale Parallel Cognitive Diagnostic Test Assembly Using A Dual-Stage Differential Evolution-Based Approach. *IEEE Transactions on Artificial Intelligence* 5, 6 (2024), 3120–3133.

[4] Hua-Hua Chang. 2015. Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1 (2015), 1–20.

[5] Hua-Hua Chang. 2015. Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1 (2015), 1–20.

[6] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics* 34, 1 (2009), 115–130.

[7] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.

[8] Z Eckart, L Marco, and T Lothar. 2001. *Improving the strength Pareto evolutionary algorithm for multiobjective optimization.* 1–21 pages.

[9] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

[10] Aritra Ghosh and Andrew S. Lan. 2021. BOBCAT: Bilevel Optimization-Based Computerized Adaptive Testing. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. Montreal, Canada, 2410–2417.

[11] David E Golberg. 1989. Genetic algorithms in search, optimization, and machine learning. *Addion wesley* 1989, 102 (1989), 36.

[12] Yuting Hong, Shiwei Tong, Wei Huang, Yan Zhuang, Qi Liu, Enhong Chen, Xin Li, and Yuanjing He. 2023. Search-Efficient Computerized Adaptive Testing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. Birmingham, United Kingdom, 773–782.

[13] Liya Hu, Zhiang Dong, Jingyuan Chen, Guifeng Wang, Zhihua Wang, Zhou Zhao, and Fei Wu. 2023. PTADisc: A Cross-Course Dataset Supporting Personalized Learning in Cold-Start Scenarios. In *Advances in Neural Information Processing Systems 36*. New Orleans, LA.

[14] Ye-shi Jiang, Ying Lin, Jing-Jing Li, Zhengjia Dai, Jun Zhang, and Xinglin Zhang. 2016. A novel genetic algorithm for constructing uniform test forms of cognitive diagnostic models. In *Proceedings of the 18th IEEE Congress on Evolutionary Computation*. British Columbia, Canada, 5195–5200.

[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceeding of the 3rd International Conference on Learning Representations*. San Diego, California.

[16] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Virtual Event, 904–913.

[17] Mingjia Li, Hong Qian, Jinglan Lv, Mengliang He, Wei Zhang, and Aimin Zhou. 2025. Foundation model enhanced derivative-free cognitive diagnosis. *Frontiers of Computer Science* 19, 1 (2025), 191318.

[18] Ying Lin, Ye-shi Jiang, Yue-Jiao Gong, Zhi-Hui Zhan, and Jun Zhang. 2019. A Discrete Multiobjective Particle Swarm Optimizer for Automated Assembly of Parallel Cognitive Diagnosis Tests. *IEEE Transactions on Cybernatics* 49, 7 (2019), 2792–2805.

[19] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. In *Advances in Neural Information Processing Systems 38*. Vancouver, Canada.

[20] Sannyuya Liu, Qing Li, Xiaoxuan Shen, Jianwen Sun, and Zongkai Yang. 2024. Automated discovery of symbolic laws governing skill acquisition from naturally occurring data. *Nature Computational Science* 4, 5 (2024), 334–345.

[21] Shuo Liu, Junhao Shen, Hong Qian, and Aimin Zhou. 2024. Inductive Cognitive Diagnosis for Fast Student Learning in Web-Based Online Intelligent Education Systems. In *Proceedings of the ACM on Web Conference 2024*. Singapore.

[22] Yuanhao Liu, Yiya You, Shuo Liu, Hong Qian, Ying Qian, and Aimin Zhou. 2025. A Fast-Adaptive Cognitive Diagnosis Framework for Computerized Adaptive Testing Systems. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*. Montreal, Canada.

[23] Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

[24] Haiping Ma, Changqian Wang, Hengshu Zhu, Shangshang Yang, Xiaoming Zhang, and Xingyi Zhang. 2024. Enhancing Cognitive Diagnosis Using Un-interacted Exercises: A Collaboration-Aware Mixed Sampling Approach. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 8877–8885.

[25] Dena F. Mujtaba and Nihar R. Mahapatra. 2021. Multi-objective optimization of item selection in computerized adaptive testing. In *Proceedings of 21th GECCO Conference on Genetic and Evolutionary Computation*. Lille,France, 1018–1026.

[26] Annibale Panichella. 2019. An adaptive evolutionary algorithm based on non-euclidean geometry for many-objective optimization. In *Proceedings of the 21st Genetic and Evolutionary Computation Conference*. Prague, Czech Republic, 595–603.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. British Columbia, Canada, 8024–8035.

[28] Yang Pian, Muyun Li, Yu Lu, and Penghe Chen. 2024. From "Giving a Fish" to "Teaching to Fish": Enhancing ITS Inner Loops with Large Language Models. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education*, Vol. 2151. Recife, Brazil, 362–368.

[29] Hong Qian, Shuo Liu, Mingjia Li, Bingdong Li, Zhi Liu, and Aimin Zhou. 2024. ORCDF: An Oversmoothing-Resistant Cognitive Diagnosis Framework for Student Learning in Online Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain, 2455–2466.

[30] Yuhu Shang, Xuexiong Luo, Lihong Wang, Hao Peng, Xiankun Zhang, Yimeng Ren, and Kun Liang. 2023. Reinforcement Learning Guided Multi-Objective Exam Paper Generation. In *Proceedings of the 2023 SIAM International Conference on Data Mining*. Paul Twin Cities, MN, 829–837.

[31] Junhao Shen, Hong Qian, Wei Zhang, and Aimin Zhou. 2024. Symbolic Cognitive Diagnosis via Hybrid Optimization for Intelligent Education Systems. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 14928–14936.

[32] Yishen Song, Liming Guo, and Qinhua Zheng. 2025. Measuring scientific inquiry ability related to hands-on practice: An automated approach based on multimodal data analysis. *Education and Information Technologies* 30, 4 (2025), 4381–4411.

[33] Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard, and Yolaine Bourda. 2017. A review of recent advances in adaptive assessment. *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning* (2017), 113–142.

[34] Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard, and Yolaine Bourda. 2017. A review of recent advances in adaptive assessment. *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning* (2017), 113–142.

[35] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY.

[36] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2023. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2023), 8312–8327.

[37] Hangyu Wang, Ting Long, Liang Yin, Weinan Zhang, Wei Xia, Qichen Hong, Dingyin Xia, Ruiming Tang, and Yong Yu. 2023. GMOCAT: A Graph-Enhanced Multi-Objective Method for Computerized Adaptive Testing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Long Beach, CA, 2279–2289.

[38] Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Jiarui Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. AI for Education (AI4EDU): Advancing Personalized Education with LLM and Adaptive Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain, 6743–6744.

[39] L Darrell Whitley et al. 1989. *The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best.* Colorado State University, Department of Computer Science.

[40] L Darrell Whitley et al. 1989. *The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best.* Colorado State University, Department of Computer Science.

[41] Zhengyang Wu, Tao He, Chenjie Mao, and Changqin Huang. 2020. Exam Paper Generation Based on Performance Prediction of Student Group. *Information Sciences* 532 (2020).

[42] Xiaoshu Xiang, Ling Wu, Haiping Ma, and Xingyi Zhang. 2023. Balancing Measurement Efficiency, Test Diversity an d Security for Item Selection in On-the-Fly Assembled Multistage Adaptive Testing via a Multi-objective Evolutionary Algorithm. In *Proceeding of the 14th International Conference on Swarm Intelligence*. Guangdong, China, 438–451.

[43] Hefei Xu, Min Hou, Le Wu, Fei Liu, Yonghui Yang, Haoyue Bai, Richang Hong, and Meng Wang. 2025. Fair Personalized Learner Modeling Without Sensitive Attributes. In *Proceedings of the ACM on Web Conference 2025*. Sydney, Australia, 4612–4624.

[44] Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, Juanzi Li, and Jie Tang. 2023. MoocRadar: A Fine-grained and Multi-aspect Knowledge Repository for Improving Cognitive Student Modeling in MOOCs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Taiwan,China, 2924–2934.

[45] Junhao Yu, Yan Zhuang, Zhenya Huang, Qi Liu, Xin Li, Rui Li, and Enhong Chen. 2024. A Unified Adaptive Testing System Enabled by Hierarchical Structure Search. In *Proceeding of the 41th International Conference on Machine Learning*. Vienna, Austria.

[46] Dacao Zhang, Kun Zhang, Le Wu, Mi Tian, Richang Hong, and Meng Wang. 2024. Path-Specific Causal Reasoning for Fairness-aware Cognitive Diagnosis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain, 4143–4154.

[47] Qingfu Zhang and Hui Li. 2007. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation* 11, 6 (2007), 712–731.

[48] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. 2022. Fully Adaptive Framework: Neural Computerized Adaptive Testing for Online Education. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. virtual event, 4734–4742.

[49] Yan Zhuang, Qi Liu, Guanhao Zhao, Zhenya Huang, Weizhe Huang, Zachary A. Pardos, Enhong Chen, Jinze Wu, and Xin Li. 2023. A Bounded Ability Estimation for Computerized Adaptive Testing. In *Advances in Neural Information Processing Systems 36*. New Orleans,LA.

APPENDIX

A Pseudocode of the Dataset Densification Algorithm

The dataset densification algorithm is introduced in Section 5.1, Algorithm 2 shows its complete procedure.

Algorithm 2 Dataset Densification

Input: Student set $\mathcal{S}$, Question bank $\mathcal{Q}$, Response logs $\mathcal{R}$
Initialization:
Transformed records $\mathcal{R}$ into a matrix $\boldsymbol{M_R} \in \{0,1\}^{n\times|\mathcal{S}|}$, where $\boldsymbol{M_R}(i,j) = 1$ means the response log of student S_i to question Q_j exist in $\mathcal{R}$, otherwise $\boldsymbol{M_R}(i,j) = 0$.
The filtered student set $\mathcal{S}_U \leftarrow \mathcal{S}$, the filtered question set $\mathcal{Q}_U \leftarrow \mathcal{Q}$
Boolean variable to indicate whether it is dense $dense \leftarrow False$
while $dense == False$ **do**
  $\boldsymbol{M}^*_{\boldsymbol{R}} \leftarrow \boldsymbol{M_R}$
  if $\exists S_i, \exists Q_j$ that $\boldsymbol{M}^*_{\boldsymbol{R}}(i,j) ==$ **empty then**
    $\boldsymbol{N_S} \leftarrow$ the number of responses for each student
    $\boldsymbol{N_Q} \leftarrow$ the number of responses for each question
    $s^* \leftarrow \arg\min(\boldsymbol{N_S})$
    $q^* \leftarrow \arg\min(\boldsymbol{N_Q})$
    $\boldsymbol{N_{s^*}} \leftarrow$ the number of responses for s^*
    $\boldsymbol{N_{q^*}} \leftarrow$ the number of responses for q^*
    if $\boldsymbol{N_{s^*}} > \boldsymbol{N^*_q}$ **then**
      $\mathcal{Q}_U \leftarrow \mathcal{Q}_U \setminus q^*$
    else
      $\mathcal{S}_U \leftarrow \mathcal{S}_U \setminus s^*$
    end if
  else
    $dense \leftarrow True$
  end if
end while
Output: The densified response logs $\mathcal{R}_D$ with the filtered student set $\mathcal{S}_U$ and the filtered question bank $\mathcal{Q}_U$

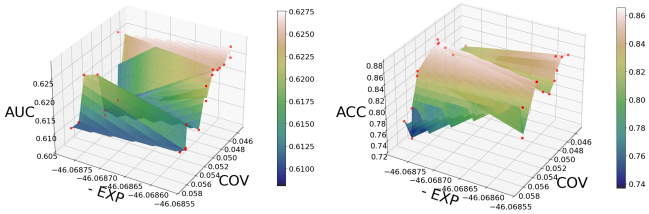

Figure 6: The Pareto front of the population assembled by PCAT on MOOC when $Step = 20$.

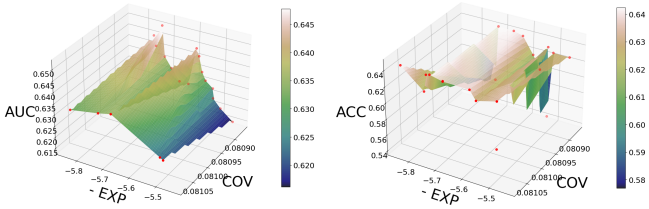

Figure 7: The Pareto front of the population assembled by PCAT on PROB-STA when $Step = 20$.

Table 4: Comparison with Existing MOO algorithms

MOO	AUC	ACC
# NSGA2	0.872	0.800
# SPEA2	0.868	0.800
# AGEMOEA	0.874	0.802

B Complete Results of Objective Comparison

The result of the objective comparison on dataset **ASSIST** and related analysis are presented in Section 5.4.4. Here are the complete results on the remaining 3 datasets as shown in Figure 6, 7 and 3.

C Analysis of Hyper-Parameters

To investigate the impact of hyper-parameters of MOO on the results of PCAT, we compare the AUC metric under different population size and the maximum number of iterations. Specifically, to study the effect of maximum iteration, we fix the population size = 100 and vary the maximum iteration within the range $\{40, 60, 80, 100, 120\}$. Similarly, to study the effect of population size, we fix the maximum iteration = 80 and vary the population size within the range $\{50, 75, 100, 125, 150\}$. The results are shown in Figure 8 and 9, respectively.

From Figure 8, which shows the impact of the maximum number of iterations on the AUC metric, it is evident that the number of iterations has minimal influence on the outcome. The AUC values remain stable across different iteration counts, indicating that the PCAT framework converges quickly and does not require a large number of iterations to achieve optimal performance. This suggests that the optimization process in PCAT is highly efficient and robust, even with a relatively small number of iterations. Similarly, Figure 9 illustrates the effect of population size on the AUC metric. The results show that varying the population size has little impact on

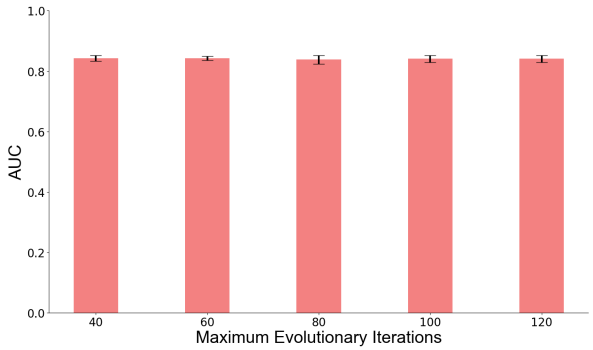

Figure 8: The AUC of PCAT w.r.t. different maximum evolutionary iterations when population size = 100.

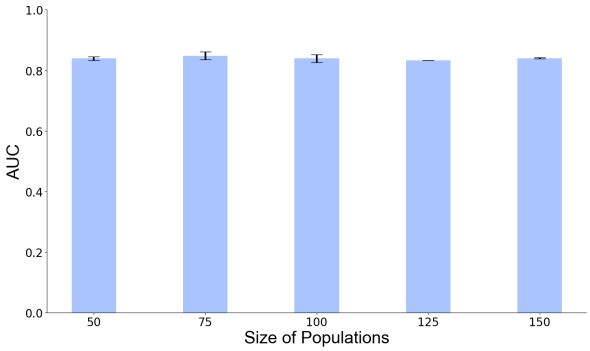

Figure 9: The AUC of PCAT w.r.t. different population size when maximum iterations = 80.

the AUC values. This indicates that the PCAT framework is not sensitive to changes in population size, further demonstrating its stability and robustness.

Meanwhile, we further investigated the impact of adopting different MOO algorithm on performance under the PCAT framework. To this end, 2 extra MOO algorithms, namely SPEA2 [8] and AGEMOEA [26] were evaluated as backbone components while maintaining identical hyperparameter configurations. Method performance was compared based on the AUC and ACC metrics at step 20, with the results presented in Table 4. As observed, the ACC remained nearly consistent across all MOO algorithms, with variations limited to less than 1% in AUC. These findings suggest that choice of MOO has minimal influence on the overall performance within the PCAT framework. This further demonstrates the robustness of our framework, as its performance remains stable regardless of the specific MOO strategy employed.

Overall, the hyper-parameter analysis reveals that both the maximum number of iterations and population size have minimal impact on the performance of the PCAT framework. This suggests that these hyper-parameters do not significantly influence or disturb the experimental results. Additionally, the selection of MOO algorithms demonstrates a a negligible influence on experimental performance metrics (<1% variation across all scenarios), as quantitatively validated in our comparative analysis. Further highlighting the strong stability and robustness of the PCAT framework in optimizing the paper assembly problem for high-stakes examinations.