Quality meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing

Haoyang Bi[1], Haiping Ma[2], Zhenya Huang[1,*], Yu Yin[1], Qi Liu[1], Enhong Chen[1], Yu Su[3], Shijin Wang[3]

[1]Anhui Province Key Laboratory of Big Data Analysis and Application
School of Computer Science and Technology, University of Science and Technology of China
{bhy0521, yxonic}@mail.ustc.edu.cn, {huangzhy, qiliuql, cheneh}@ustc.edu.cn
[2]Anhui University, hpma@ahu.edu.cn
[3]IFLYTEK Research, {yusu, sjwang}@iflytek.com

Abstract—Computerized Adaptive Testing (CAT) is emerging as a promising testing application in many scenarios, such as education, game and recruitment, which targets at diagnosing the knowledge mastery levels of examinees on required concepts. It shows the advantage of tailoring a personalized testing procedure for each examinee, which selects questions step by step, depending on her performance. While there are many efforts on developing CAT systems, existing solutions generally follow an inflexible *model-specific* fashion. That is, they need to observe a specific cognitive model which can estimate examinee's knowledge levels and design the selection strategy according to the model estimation. In this paper, we study a novel *model-agnostic* CAT problem, where we aim to propose a flexible framework that can adapt to different cognitive models. Meanwhile, this work also figures out CAT solution with addressing the problem of how to generate both high-quality and diverse questions simultaneously, which can give a comprehensive knowledge diagnosis for each examinee. Inspired by Active Learning, we propose a novel framework, namely Model-Agnostic Adaptive Testing (MAAT) for CAT solution, where we design three sophisticated modules including Quality Module, Diversity Module and Importance Module. Specifically, at one CAT selection step, Quality Module first quantifies the informativeness of questions and generates candidate subset with the highest quality. Then, Diversity Module selects one question at each step that maximizes the concept coverage. Additionally, we propose Importance Module to estimate the importance of concepts that optimizes the CAT selection. Under MAAT, we prove that the goal of maximizing both quality and diversity is NP-hard, but we provide efficient algorithms by exploiting the inherent submodular property. Extensive experimental results on two real-world datasets clearly demonstrate that our MAAT can support CAT with guaranteeing both quality and diversity perspectives.

Index Terms—Computerized Adaptive Testing, Model-Agnostic, Quality, Diversity

I. INTRODUCTION

Designing appropriate tests to evaluate the knowledge states on required concepts of examinees is a fundamental task in many real-world scenarios, such as education, game and job recruit [1], [2]. Traditionally, instructors can organize a pencil-paper test, which carefully selects a set of questions for examinees at one time, and therefore we can assess the states of them from their performances. Although such simple way is

effective, it just provides all examinees with the same environment so it is difficult to guarantee the rationality of all selected questions [3]. Therefore, recent efforts focus on another testing form called *Computerized Adaptive Testing* (CAT), which aims to build tests that personally adapt to each examinee, tailoring questions step by step, depending on her performances [4]. In fact, CAT has many advantages including improving accuracy, guaranteeing security and enhancing examinee engagement, which has already been applied in many standard test organizations, such as Graduate Management Admission Test (GMAT) [5] and Graduate Record Examinations (GRE) [6].

In practice, a typical CAT system generally consists of two key components [3], [7]: (1) a *cognitive diagnosis model* (CDM) that estimates the knowledge states of examinees according to their performance; (2) a *selection strategy* that chooses a question from the pool to support the testing procedure. As shown in Fig. 1(a), when an examinee e_1 comes, such CAT system can establish an interactive testing procedure for her. At step t, the system first posts one question (e.g., q_t). Then, she reads and answers it. After receiving the response (i.e., right or wrong), the system with CDM estimates her current states and on the basis carefully selects a new question q_{t+1} at the next round. This procedure repeats several times until meeting the termination like reaching the maximum testing length [8], so that we can realize how much she has learned about the required concepts (e.g., "Function" in Math). In this way, even if starting with the same question, examinees, e.g., e_1 and e_2 in Fig. 1(b), still can be tailored personalizations. Therefore, the key issue is how to establish an optimal CAT system for choosing the appropriate questions for examinees.

In the literature, there are many efforts on designing CAT, which have already supported many standard tests [9]. Generally, existing solutions deeply dig into underlying CDMs, such as item response theory (IRT) [10] and multidimensional one (MIRT) [11], and then produces questions via observing the corresponding model parameters related to examinees' knowledge states. For example, Lord et al. [12] established a CAT system, where they proposed a maximum fisher information strategy that greedily selected the questions with minimizing the variance of examinee's parameters obtained by specific IRT
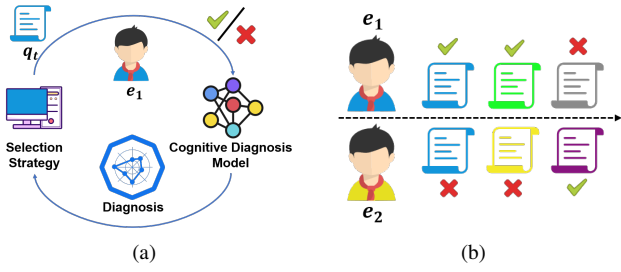
Fig. 1. (a) Illustration of a typical CAT system and its testing procedure in one step. (b) Two toy examples of CAT procedure taken by examinees e_1 and e_2. We represent different questions with different colors.

model. Although these works have made great success, they are all **model-specific**, i.e., the CAT solution is only suitable for its designated CDM. In other words, we have to understand how a specific CDM (e.g., IRT or MIRT) works in detail when designing a CAT selection strategy. Therefore, existing CAT systems usually become inflexible since we must redesign the selection strategy if we replace the CDM behind. However, as we mentioned above, CAT primarily aims to select appropriate questions during testing so it gives us an intuition that we only need to care about the results of examinees' states no matter which CDM models we probe. To this end, we argue that an ideal CAT framework should be **model-agnostic**, i.e., the CAT solution can adapt to different CDMs.

To the best of our knowledge, no previous work on CAT has attempted to achieve the model-agnostic framework. Fortunately, we notice a similar idea from active learning (AL), which is a popular framework to alleviate data deficiency in many tasks including image classification, recommendation, etc. [13], [14]. In principle, AL framework aims to design a strategy that selects valuable data step by step to experts for annotations so that machine learning models can be well trained in the supervised manner. Intuitively, AL inspires us for the CAT solution since it can also overlook which machine learning models we have to use. Therefore, in this paper, we propose a novel **M**odel-**A**gnostic **A**daptive **T**esting framework (MAAT) for CAT solution, where we take advantage of the general idea of active learning at the macro level.

In order to support MAAT procedure that evaluates the knowledge states of examinees comprehensively, we argue that there are two necessary objectives should be considered in designing the strategy for adaptively selecting questions: (1) *Quality*. Primarily, selecting high-quality questions for testing can make the CAT procedure more efficiently. For example, it is inappropriate if we always push examinees to try questions that are either too difficulty or too easy, because we cannot obtain the accurate diagnosis results about them at all. Therefore, an effective method that can evaluate the informativeness of questions is urgent but non-trivial. (2) *Diversity*. In a certain domain (e.g., math), there is usually much knowledge (e.g., "Function") that examinees should learn. With CAT process, we are required to evaluate that how they master all the knowledge concepts. However, if we follow the traditional solutions [8], [9], the results may be suboptimal, since few of them directly consider such diversity

issue, leading to very limited question selections on concepts. Therefore, it is necessary to select questions in MAAT that cover concepts as much as possible.

To address the above problem with considering both objectives above, we implement our MAAT framework with proposing three modules, i.e., Quality Module, Diversity Module and Importance Module. Specifically, at one CAT selection step, Quality Module first generates a small candidate subset of the most high-quality questions from the pool, where a novel score function is proposed to quantify the information gain of questions on knowledge mastery after the examinee has taken. Different from existing solutions, this module is flexible since it evaluates the informativeness through the Expected Model Change (EMC) without awareness of the detailed mechanism behind the change. Then, Diversity Module selects one question from candidates that maximizes the concept coverage in the whole CAT procedures. We also propose Importance Module to estimate the importance of concepts that optimizes the selection procedure. Moreover, we prove that the problem of maximizing both quality and diversity becomes NP-hard under our sophisticated coverage score function and provide efficient algorithms by exploiting the submodular property. Finally, we conduct extensive experiments on two real-world datasets. The experimental results demonstrate that our MAAT can select both high-quality and diverse questions in a model-agnostic way, which can support many CAT scenarios.

II. RELATED WORK

1) Computerized Adaptive Testing: The development of Computerized Adaptive Testing (CAT) originates from the belief that tests can be more effective if we tailor them for examinees [4]. Though there are some variants recently [15], [16], the primary challenge of CAT lies in designing a selection strategy which selects appropriate questions for the examinee step by step. Since current strategies are closely bound to the underlying cognitive diagnosis models (CDM) [17], [18], we review them in terms of the CDMs they base on. Representative CDMs include traditional item response theory (IRT) family [10], [11] and recently proposed deep learning models [19], [20]. IRT-based CAT strategies minimize the statistical estimation error of the latent parameter in IRT [9], [21]. MIRT-based strategies are proposed as multivariate extensions of the IRT-based ones [22]–[24]. To the best of our knowledge, little progress has been made in designing strategies for the deep learning models due to their parametric complexity. Having made great effect though, current strategies suffer from two limitations. First, they must understand how the underlying CDM works in detail, making CAT systems model-agnostic and inflexible. Second, they overlook the diversity in question selection, causing potential imbalance in the diagnosis for knowledge concept mastery. We provide novel solutions within our proposed model-agnostic framework, which will be discussed in Section IV.

2) Active Learning: Active learning (AL) is motivated by the belief that we can train a better model with less data if we actively select valuable data, and has been applied in many

supervised learning tasks [13], [14]. Starting with a machine learning model and a data selection strategy, at each step, AL framework selects a batch of unlabeled data to be annotated for supplementing the limited labeled data so that the model achieves better performance. The key point is how to avoid using model details in strategy design so that it can apply to varieties of tasks with different models. Generally, there are two solutions which utilize model outputs and data features, respectively [25]. Specifically, *uncertainty* based algorithms examine the label predictions output by the model and select the ones whose predictions contain the most uncertainty [13], [26]; *representativeness* based algorithms examine the features of data samples and select the ones which represent the overall patterns of unlabeled data best [27]. The methodologies inspire us to propose a model-agnostic solution for CAT as well, however with different goals dedicated to CAT (i.e., quality and diversity) and a novel strategy to optimize the goals.

3) Coverage Measure: The coverage measure has been extensively studied in tasks related to document summarization [28], [29], network analysis [30] and recommendation [31], [32], in which we try to find a subset that covers as much information in the document or recommendation as possible. In many scenarios such as recommendation, it is intuitively better to consider such coverage objective [31], [33]–[36]. Specially, some previous works utilize *submodularity* in their design of coverage score function for optimization [28], [30], [31], which is a mathematical modeling towards the intuitive diminishing returns property. To the best of our knowledge, our work is the first attempt to explicitly define the diversity goal of knowledge concepts in CAT with a formulated coverage measure, where we provide an efficient optimization algorithm by exploiting the inherent submodularity property.

III. PRELIMINARIES

This section discusses the terminologies, the goals, and the reformulation of computerized adaptive testing (CAT).

A. Terminologies

1) Environment: As a specific form of test, a CAT system works with a typical testing environment consisting of examinees and questions. Suppose there is an examinee set $E = \{e_1, e_2, ..., e_{|E|}\}$ and a question set $Q = \{q_1, q_2, ..., q_{|Q|}\}$. We denote the record of examinee e_i answering question q_j as a triplet $r_{ij} = <e_i, q_j, a_{ij}>$, where a_{ij} equals 1 if e_i answers q_j correctly, and 0 otherwise. We denote all the records as R, and the records belonging to a certain examinee e_i as R_i. In addition, we suppose a set of knowledge concepts $K = \{k_1, k_2, ..., k_{|K|}\}$ related to the questions. We denote the association between questions and concepts as a binary relation $G \subseteq Q \times K$, where $(q_i, k_j) \in G$ if q_i is related to k_j.

2) Status: Besides the static environment, a CAT system maintains some dynamic status dedicated to adaptive tests. Specifically, within the test for a certain examinee $e_i \in E$, the question set Q is divided into a tested set Q_T and an untested set Q_U. Initially, $Q_U = Q, Q_T = \emptyset$. At each step, one question is selected from Q_U to Q_T. When the test finishes, Q_T forms a tailored test sequence for e_i.

TABLE I
CONCEPT CORRESPONDENCE BETWEEN CAT AND AL

CAT Concepts	AL Concepts	Notation
Cognitive diagnosis model	Supervised learning model	$\mathcal{M}$
Question selection strategy	Sample query strategy	$\mathcal{S}$
Examinees	Expert annotators	E
Questions	Data samples	Q
Tested/Untested questions	Labeled/Unlabeled samples	Q_T, Q_U

3) Components: Finally, we formulate the components of a CAT system. Following Fig. 1(a), we formally denote a CAT system as $(\mathcal{M}, \mathcal{S})$, where $\mathcal{M}$ is a cognitive diagnosis model (CDM) and $\mathcal{S}$ is a question selection strategy. Different from traditional CAT systems, in our problem, $\mathcal{M}$ does not refer to any specific CDM (e.g., IRT), but an abstract model with two basic functionalities: (1) $\mathcal{M}$ captures the knowledge states of the examinees with a group of parameters $\boldsymbol{\theta}$ without any assumption about the detailed form or mechanism; (2) given an examinee $e_i \in E$ and a question $q_j \in Q$, $\mathcal{M}$ can output a performance prediction $\mathcal{M}(e_i, q_j|\boldsymbol{\theta}) \in [0, 1]$ which measures how likely e_i can answer q_j correctly. $\mathcal{S}$ accepts Q_U and $\mathcal{M}$ as input, and outputs a question $q \in Q_U$, i.e., $q = \mathcal{S}(Q_U, \mathcal{M})$. In other words, it makes the selection from the untested question set according to the current estimated knowledge states.

B. Goals

We now discuss the two goals for the selection strategy $\mathcal{S}$.
1) Quality: Generally, a high-quality question helps reduce the uncertainty of the examinee's knowledge states. Therefore, we quantify the quality of a question through its *informativeness*, i.e., how much information the underlying model $\mathcal{M}$ can obtain from the question to update the estimate for knowledge states. In this way, achieving the quality goal means to select the most informative questions. To evaluate informativeness, after the test for $e_i \in E$, we predict the her performance with $\mathcal{M}$ on the whole question pool, and measure it with some metric such as AUC. We denote such measurement as $Inf(\mathcal{S})$, which will be discussed in detail in Section V.

2) Diversity: Generally, we consider a set of questions to be diverse if it meets certain coverage requirements. We intuitively measure diversity with knowledge coverage. As a result, achieving the diversity goal means to select a set of questions that has the maximum knowledge concept coverage. After the test, We can evaluate the coverage by the proportion of the knowledge concepts the tested question set Q_T covers, denoted as $Cov(\mathcal{S})$. We will discuss it in detail in Section V.

C. Problem Formulation

Inspired by active learning (Table I), we reformulate our model-agnostic CAT problem, however, with a key difference that we aim to achieve the problem considering both quality and diversity goals as discussed above:

Problem Definition. Given a new examinee $e_i \in E$, a question pool Q with knowledge concepts K, our task is to design a strategy $\mathcal{S}$ to select a N-size question set $Q_T = \{q_1^*, q_2^*, ..., q_N^*\}$ step by step that has the maximum quality and diversity. Before testing, we set up an abstract CDM $\mathcal{M}$ with parameters $\boldsymbol{\theta}$ capturing knowledge states. During testing,
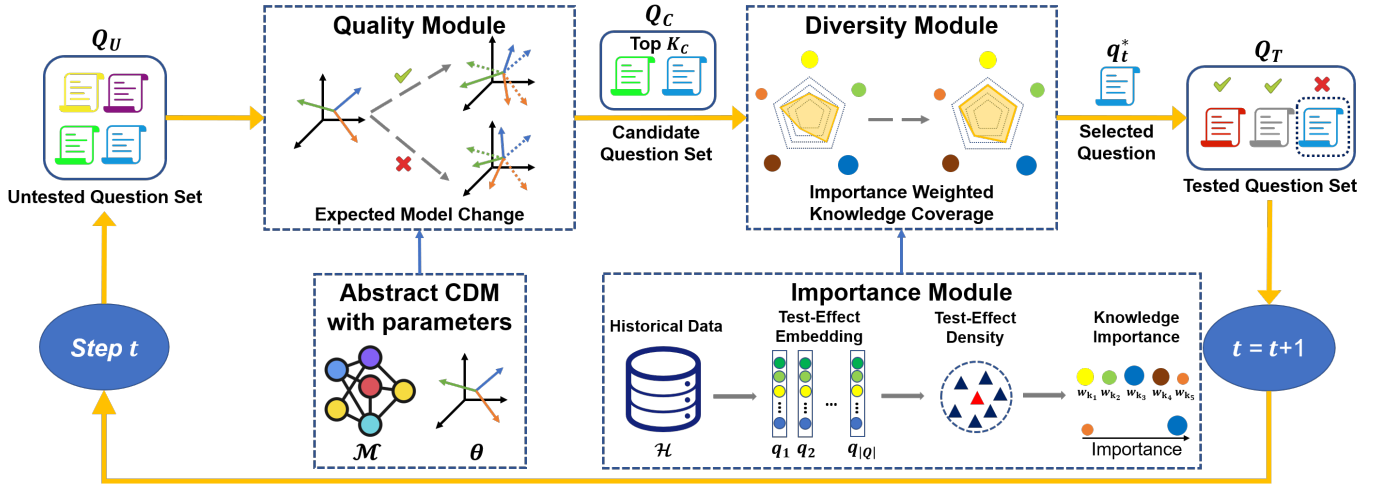
Fig. 2. Overview of our MAAT framework.

at step t ($1 \le t \le N$), we select one question $q_t^* = \mathcal{S}(Q_U, \mathcal{M})$, then observe a new interaction test record $r_{it}^* = < e_i, q_t^*, a_{it}^* >$ and update the knowledge states, i.e., $\boldsymbol{\theta}$, in $\mathcal{M}$ instantly. After testing, we measure the effectiveness of $\mathcal{S}$ by computing $Inf(\mathcal{S})$ and $Cov(\mathcal{S})$.

IV. MAAT FRAMEWORK

A. Overview

We present the overview of our MAAT framework in Fig. 2. For any examinee $e_i \in E$, the proposed framework appropriately generates a question set step by step in interaction with her. At step t, MAAT selects one question from the untested question set Q_U to the tested quesion set Q_T, which optimizes both quality and diversity goals. The overall architecture can be seen as three modules: *Quality Module*, *Diversity Module* and *Importance Module*. Specifically, at step t, in Quality Module, a candidate question set Q_C is selected from Q_U, which consists of the top-K_C high-quality questions with maximum informativeness. Then in Diversity Module, one final question q_t^* is selected from Q_C to Q_T, which contributes the maximum marginal gain to the knowledge coverage of Q_T. Additionally, to achieve a more efficient selection procedure, we propose Importance Module to evaluate the importance of knowledge concepts, which utilizes historical data for pre-computation before testing.

B. Quality Module

We first introduce Quality Module, which is the first stage of step t as shown in the upper left part of Fig. 2. The aim of this module is to select a candidate set Q_C from Q_U consisting of the top-K_C high-quality questions ($K_C = |Q_C|$). To achieve that, we propose a score function, namely *Expected Model Change* (EMC), which quantifies the quality of questions by measuring their informativeness, i.e., how much information they contain. Different from traditional CAT heuristic functions [8], [9], EMC is independent of the details of the CDM.

The general idea behind is as follows. We make use of the information contained in a question by updating our estimate to the examinee's knowledge states after observing her answer. Therefore, the informativeness can be scored by how much the diagnosis changes through a question. In MAAT, the knowledge states are carried by the parameters of the abstract CDM, i.e., $\boldsymbol{\theta}$ in $\mathcal{M}$. Regardless of the concrete mechanism, how much the CDM changes implies the amount of information obtained from the question. Specifically, if $\boldsymbol{\theta}$ has a great change, the question can be considered to be informative. Otherwise, if $\boldsymbol{\theta}$ hardly changes, knowing the answer to the question brings little information.

The challenge is that it is impossible to know the examinee's response before selection. Therefore, we compute the expectation of model change w.r.t. the probability that the examinee answers the question correctly, which is predicted by $\mathcal{M}$. Formally, let $\Delta\mathcal{M}(r_{ij}) = |\boldsymbol{\theta}(R_i \cup \{r_{ij}\}) - \boldsymbol{\theta}(R_i)|$ be the model change that would be obtained by adding the record $r_{ij} = < e_i, q_j, a_{ij} >$, where $\boldsymbol{\theta}(R_i)$ denotes the parameters trained with e_i's current records, R_i, and so as $\boldsymbol{\theta}(R_i \cup \{r_{ij}\})$. For every $q_j \in Q_U$, we define its *expected model change* as:

$$\text{EMC}(q_j) = \mathbb{E}_{a_{ij} \sim p} \Delta\mathcal{M}(< e_i, q_j, a_{ij} >), \quad (1)$$

$$p = \mathcal{M}(e_i, q_j | \boldsymbol{\theta}(R_i)). \quad (2)$$

For computational efficiency, we approximate $\Delta\mathcal{M}(r_{ij})$ with the gradient caused by r_{ij} instead of retraining the model. This approaximation is especially efficient for those CDMs using gradient-based training, such as neural models.

With the EMC score function, for each untested question $q_j \in Q_U$, we evaluate $\text{EMC}(q_j)$ (Eq. (1)), and then select a candidate set of the top-K_C high-quality questions Q_C which have the maximum informativeness.

C. Diversity Module

Now that we have a candidate set Q_C containing high-quality questions, we turn to Diversity Module, which is the second stage of step t as shown in the upper right part of Fig. 2. In this module, we aim at selecting one question q_t^* from Q_C that optimizes the diversity goal. The neccessity of this stage comes from the observation that our diagnosis will be likely

one-sided if we overlook the diversity in question selection. To achieve high diversity as well, we first propose a score function that quantifies the knowledge coverage of the tested question set Q_T, and then search for an algorithm to construct Q_T with maximum coverage score by adding questions step by step. The two main challenges are how to properly construct the coverage score function and how to design the optimization algorithm.

To solve the first challenge, we begin with a Naive Knowledge Coverage (NKC) function, which simply calculates the proportion of knowledge concepts covered by the selected question set Q_T to all the knowledge concepts:

$$\text{NKC}(Q_T) = \frac{\sum_{k \in K} Cov(k, Q_T)}{|K|}, \quad (3)$$

$$Cov(k, Q_T) = \mathbb{1}[\exists q \in Q_T, (q, k) \in G]. \quad (4)$$

Although NKC is intuitive, we find it suffer from two practical drawbacks: (1) It treats all the knowledge concepts equally and cannot distinguish their importance. For example, in a math test, if our tests focus more on "Algebra" than "Geometry", we should select more questions related to "Algebra" besides covering both of them. (2) It is too strict since $Cov(k, Q_T)$ is binary depending on whether k is covered by Q_T, which means that $Cov(k, Q_T)$ always equals to 1 as long as at least one question in Q_T is related to k, regardless of the exact count. This strong condition might result in the imbalance of covered knowledge. For example, choosing 9 questions related to "Algebra" and 1 question related to "Geometry" is equivalent to choosing 5 of each under this condition, however the latter is usually a better choice.

To this end, following the idea in [31], we add two novel features to NKC (Eq. (3)): (1) to take importance of knowledge concepts into consideration, we add an importance weight w_k for each knowledge concept $k \in K$; (2) to alleviate the imbalance caused by the strict binary $Cov(k, Q_T)$ in Eq. (4), we improve it into a soft form with incremental property, i.e., the value gradually grows from 0 to 1 as there are more and more questions related to the knowledge concept. Formally, we proposed an advanced score function named *Importance Weighted Knowledge Coverage* (IWKC):

$$\text{IWKC}(Q_T) = \frac{\sum_{k \in K} w_k \times IncCov(k, Q_T)}{\sum_{k \in K} w_k}, \quad (5)$$

$$IncCov(k, Q_T) = \frac{cnt(k, Q_T)}{cnt(k, Q_T) + 1}, \quad (6)$$

$$cnt(k, Q_T) = \sum_{q \in Q_T} \mathbb{1}[(q, k) \in G], \quad (7)$$

where w_k is the added importance weight for concept k, which is a positive constant. We will discuss the computation of w_k in Section IV-D. $IncCov(k, Q_T)$ is the incremental improvement to $Cov(k, Q_T)$ in Eq. (4). For example, when the number of questions related to k is 0, 1, 2, 3, ..., $IncCov(k, Q_T)$ gradually reaches 0, 0.5, 0.67, 075, ..., respectively, while as $Cov(k, Q_T)$ discontinuously jumps from 0 to 1.

Note that though IWKC has been well constructed, the sophisticated structure still brings the second challenge, i.e.,

how to select a set of questions which has the maximum IWKC score in a step-by-step way. Indeed, this optimization problem is proved to be NP-hard (Section IV-E). Fortunately, we find a suboptimal solution with acceptable performance by exploiting the *submodular* property of IWKC. Generally, submodularity can be seen as a mathematical modelling to the narural *diminishing returns* property [37]. For a submodular set function, as the set gets larger, the marginal gain obtained by adding one more element will decrease. Specifically, in our case, as Q_T grows larger with selection steps going on, the gain in coverage (i.e., IWKC) caused by adding the same question will get slower. For clarity of our discussion, we leave the formal proof later in Section IV-E. The submodular property of IWKC provides us with a performance-guaranteed greedy selection algorithm [37]. Generally, at step t in the test, for each candidate question $q_j \in Q_C$, we evaluate the marginal gain of IWKC for Q_T if q_j were added to Q_T, and greedily select the one maximizing the marginal gain as the t-step selection q_t^*:

$$q_t^* = argmax_{q_j \in Q_C} \Delta_{q_j} \text{IWKC}(Q_T), \quad (8)$$

$$\Delta_{q_j} \text{IWKC}(Q_T) = \text{IWKC}(Q_T \cup \{q_j\}) - \text{IWKC}(Q_T). \quad (9)$$

With the above algorithm, the ratio of Q_T's IWKC to the optimal value is guaranteed to be at least $1 - \frac{1}{e}$.

D. Importance Module

After demonstrating the two-stage procedure to select an optimal question at each step during testing, we turn to solve the problem of computing the importance weight w_k in IWKC (Eq. (5)). As shown in the lower right part of Fig. 2, Importance Module pre-computes w_k for each knowledge concept before the test begins. The general idea is to consider a knowledge concept to be important if its associated questions are more *representative*. Typically, a question is considered to be representative if it has similar characteristics with many other questions. With a representative question, we can implicitly examine many questions at the same time. To quantify the representativeness of questions, we firstly represent them with feature vectors so that each question can be seen as a point in the embedding metric space. The closer a question is to its neighbors, the more representative the question. Finally, we obtain the importance weight of each knowledge concept by averaging the representativenss (i.e., the density) of its related questions. To accomplish the computation, we utilize the historical data of the CAT system. Specifically, we have historical examinees whose records are persisted and can be used for training, which we denote as $\mathcal{H} = (E^H, R^H)$.

1) Test-Effect Embedding: The key point of embedding questions is to define the distance metric between questions. The general idea is that the historical examinees' performance on the question can characterize the question itself, such as difficulty and differentiation. Thus we declare that questions on which examinees perform similarly have *Test-Effect similarity*, and define the question embedding following such similarity as *Test-Effect embedding*. In order to train Test-Effect embeddings, we extend the idea from *Item2Vec* [38].

Specifically, for each historical record $r_{ij} = <e_i, q_j, a_{ij}>$, we set up an input $\boldsymbol{x}_{ij}$ to represent both which question was answered and if the question was answered correctly:

$$\boldsymbol{x}_{ij} = \begin{cases} \mathbf{1}_{|Q|}(j) \oplus \mathbf{1}_{|Q|}(j), & \text{if } a_{ij} = 1 \\ \mathbf{1}_{|Q|}(j) \oplus \mathbf{0}_{|Q|}, & \text{if } a_{ij} = 0 \end{cases}, \quad (10)$$

where $\mathbf{1}_{|Q|}(j)$ denotes a $|Q|$-length one-hot vector with only the jth position equal to 1, $\mathbf{0}_{|Q|}$ denotes a $|Q|$-length zero vector, and $\oplus$ denotes vector concatenation. Then we train a *Skip-Gram Negative Sampling* (SGNS) model [39]. Specifically, given a historical examinee $e_k \in E^H$, the optimization objective is formulated as:

$$\max \frac{1}{|R_k^H|} \sum_{r_{ki} \in R_k^H} \sum_{r_{kj} \in R_k^H, j \neq i} \log p(r_{kj}|r_{ki}), \quad (11)$$

$$p(r_{kj}|r_{ki}) = \sigma((\boldsymbol{W}\boldsymbol{x}_{ki})^T \boldsymbol{v}_j) \prod_{l=1}^{N_{neg}} \sigma(-(\boldsymbol{W}\boldsymbol{x}_{kn_l})^T \boldsymbol{v}_{n_l}), \quad (12)$$

where N_{neg} is the negative sampling size, $\sigma(x) = \frac{1}{1+e^{-x}}$ is the Sigmoid function, $\boldsymbol{W}$ is a $d \times 2|Q|$ parameter matrix, and $\boldsymbol{v}_j$ is our Test-Effect embedding of q_j with dimension d, which we denote as $\boldsymbol{q}_j^{TE}$.

2) Test-Effect Density: Since the questions have been represented in the Test-Effect embedding space, we quantify the representativeness of questions. First, We compute Test-Effect similarity between two questions q_i and q_j as

$$Sim_{TE}(q_i, q_j) = e^{-\gamma |\boldsymbol{q}_i^{TE} - \boldsymbol{q}_j^{TE}|}, \quad (13)$$

where $|\cdot|$ is Euclidean norm and γ is a positive smoothing parameter. Next, we define the *Test-Effect density* of each question q_j as the average similarity to its neighbors:

$$Den_{TE}(q_j) = \frac{1}{K_N} \sum_{q_i \in \mathcal{N}(q_j)} Sim_{TE}(q_i, q_j), \quad (14)$$

where $\mathcal{N}(q_j)$ contains the K_N-nearest-neighbors of q_j in the Test-Effect embedding space. The larger Test-Effect density a question has, the more representative it is.

3) Knowledge Importance: At last, we define the knowledge importance of each knowledge concept k, i.e., w_k, by averaging the Test-Effect density of the questions related to it:

$$w_k = \frac{1}{\sum_{(q,k) \in G} 1} \sum_{(q,k) \in G} Den_{TE}(q). \quad (15)$$

Substituting Eq. (15) into Eq. (5), we get the complete IWKC.

E. Theoretical Analysis

To be rigorous, we supplement the theoretical proofs about the optimization with IWKC score function (Section IV-C). Formally, we define the *IWKC Maximization Problem* as follows: given a question set Q with associated knowledge concepts K, the target is to identify a N-size subset Q_T with the maximum IWKC:

$$\max_{Q_T \subset Q, |Q_T| = N} \text{IWKC}(Q_T). \quad (16)$$

First, we demonstrate the complexity of this problem:

Theorem. *The IWKC maximization problem is NP-hard.*

Proof: First, we introduce a classic NP-hard question, namely *weighted maximum coverage problem*: given a number

k, a collection of sets $S = \{S_1, S_2, ..., S_m\}$, a domain of elements $E = \{e_1, e_2, ..., e_n\}$ each of which has a weight $w_i, i = 1, ..., n$, the objective is to find a subset $S_{opt} \subset S$ such that $|S_{opt}| \leq k$ and the total weights of the elements covered by $\bigcup_{S_i \in S_{opt}} S_i$ is maximized. Next, we reduce the IWKC maximization problem to the weighted maximum coverage problem. We consider the knowledge concepts as the corresponding elements, and the question set $Q = \{q_1, q_2, ..., q_m\}$ covering concepts as the corresponding collection of sets. The weight of each concept k corresponds to $w_k \times IncCov(k, Q)$ as defined in Eq. (5). Under this situation, the IWKC maximization problem is equivalent to the weighted maximum coverage problem, and therefore is NP-hard. ∎

Then we verify the submodular property of IWKC:

Theorem. *IWKC is a nonnegative monotone submodular function.*

Proof: The nonnegativity and monotonicity of IWKC are obvious since it grows from 0 to 1, so we focus on proving the submodularity. First, we claim that $IncCov(k, Q_T)$ (Eq. (6)) is submodular. Let $\Delta_{q_j} IncCov(k, Q_T) = IncCov(k, Q_T \cup \{q_j\}) - IncCov(k, Q_T)$ be the marginal gain of $IncCov(k, Q_T)$ when q_j is added, which is calculated as $\Delta_{q_j} IncCov(k, Q_T) = \frac{cnt(k, Q_T \cup \{q_j\}) - cnt(k, Q_T)}{(cnt(k, Q_T \cup \{q_j\}) + 1)(cnt(k, Q_T) + 1)}$. Consider the tested question sets at two consecutive steps, $Q_T' \subset Q_T'' \subset Q$. As $Q_T' \subset Q_T''$, we have $cnt(k, Q_T' \cup \{q_j\}) - cnt(k, Q_T') \geq cnt(k, Q_T'' \cup \{q_j\}) - cnt(k, Q_T'')$ and $(cnt(k, Q_T' \cup \{q_j\}) + 1)(cnt(k, Q_T') + 1) \leq (cnt(k, Q_T'' \cup \{q_j\}) + 1)(cnt(k, Q_T'') + 1)$. Thus $\Delta_{q_j} IncCov(k, Q_T') \geq \Delta_{q_j} IncCov(k, Q_T'')$, which is the definition of submodularity. Since IWKC is nonnegative linear combinations of $IncCov$, the submodularity of IWKC can be easily derived from the submodularity of $IncCov$. ∎

Finally, we review the performance guarantee of the marginal gain based greedy algorithm [37]:

Theorem. *For any nonnegative monotone submodular function F, let S^* be the N element set with the best performance and S the same size set obtained by greedy algorithm, which selects an element with maximum marginal gain each time, and then $F(S) \geq (1 - \frac{1}{e})F(S^*)$.*

In our case, IWKC corresponds to the F above, and the selected question set Q_T corresponds to the S.

F. Summary

In summary, the flowchart of MAAT framework is as Algorithm 1, which adpots a two-stage solution for the multi-objective optimization [34], [35]. MAAT has the following advantages: (1) it is model-agnostic and suitable for a wide range of CDMs, including those which traditional CAT methods cannot fit in; (2) it optimizes both quality and diversity with novel score functions and has an efficient performance-guaranteed optimization algorithm. It is worth noting that MAAT keeps a balance of the two goals with the hyperparameter K_C, the size of the candidate set connecting Quality Module and Diversity Module, which will be explored further in Section V.

Algorithm 1 MAAT Flowchart

Data: $Q, K, G, \mathcal{H}$
Input: $\mathcal{M}; e_i \in E$
Output: A N-length test sequence $Q_T = \{q_1^*, q_2^*, ..., q_N^*\}$
Initialization:
    create $\mathcal{M}$ with $\mathcal{H}$ and randomly initialize $\boldsymbol{\theta}$
    train Test-Effect embedding for $q \in Q$ (Eq. (11) (12))
    compute importance weight for $k \in K$ (Eq. (15))
    $Q_U = Q, Q_T = \emptyset$
for $t = 1$ **to** N **do**
    $\forall q_j \in Q_U$, calculate $\text{EMC}(q_j)$ (Eq. (1))
    $Q_C = \{$top K_C questions with maximum EMC$\}$
    $q_t^* = argmax_{q_j \in Q_C} \Delta_{q_j} \text{IWKC}(Q_T)$ (Eq. (9) (5))
    $Q_U = Q_U - \{q_t^*\}$
    $Q_T = Q_T \cup \{q_t^*\}$
    observe $r_{it}^* = <e_i, q_t^*, a_{it}^*>$
    update $\mathcal{M}(\boldsymbol{\theta})$
end for

V. EXPERIMENT

In this section, we evaluate our MAAT framework on two real-world datasets. In addition, we conduct an ablation study on how MAAT keeps a balance of quality and diversity. Our codes are available in https://github.com/bigdata-ustc/MAAT.

A. Dataset Description

We use two real-world educational datasets, namely EXAM and ASSIST. The EXAM dataset was supplied by iFLYTEK Co., Ltd., collected from an online educational system where students took exams for testing. On this system, we collected the data records of junior high school students on math tests as well as the associated knowledge concepts of those questions, such as "Algebra". ASSIST is an open dataset, namely *Assistments 2009-2010 skill builder**, that recorded students' practice on math. Each question contained in ASSIST is associated with one or more knowledge concepts such as "Absolute Value".

B. Experimental Setup

1) Data Preprocessing and Partition: For the sake of the reliability the experimental results, we apply the following data preprocessing. First, in both EXAM and ASSIST, we filter the knowledge concepts that have less than 10 related questions; Second, in ASSIST, we filter the questions that are answered by less than 50 students and the students that answer less than 10 questions. Detailed statistics are presented in Table II.

In our experiment, we partition the data into *historical data* and *testing data* for different purposes. The historical data, denoted as $\mathcal{H} = (E^H, R^H)$ before, are assumed known before the tests begin. $\mathcal{H}$ is used for the CDMs to initially learn some parameters fixed during testing, such as the difficulty of the questions. In addition, MAAT utilizes $\mathcal{H}$ as input for question embedding to compute the knowledge importance weights

*https://sites.google.com/site/assistmentsdata/

TABLE II
STATISTICS OF THE DATASETS

Dataset	EXAM	ASSIST
Num. students	4,307	1,505
Num. questions	527	932
Num. concepts	31	22
Num. records	105,586	59,500
Avg. records per student	24.5	39.5
Avg. records per question	200.4	63.8
Avg. questions per concept	17.0	44.38

(Section IV-D). The testing data are used to simulate an adaptive testing environment. The students in testing data are treated as examinees that are new to the CAT system, and their records are assumed unknown until we select the questions for them during testing. On the other hand, for evaluation, we limit our selection to those questions whose response has been recorded in the testing data during our experiment. Therefore, to ensure that the candidate question set is large enough, we partition those students with more records into the testing data. Specifically, for EXAM, we divide the students who answered at least 100 questions into the testing data; for ASSIST, we divide the students who answered at least 150 questions into the testing data. The remaining parts are left as historical data.

2) Parameter Setting: We set the test length $N = 50$, which is quite enough for typical tests in practical. In Quality Module, we set the size of the output candidate set, $K_C = 10$. In Importance Module, we set $N_{neg} = 10$ (Eq. (12)), $\gamma = 0.1$ (Eq. (13)), $K_N = 10$ (Eq. (14)).

C. Baseline Approaches

To evaluate our model-agnostic framework, we compare it with classic model-specific methods designated to two different CDMs. Additionally, we conduct experiments with a deep learning CDM that classic approaches do not fit in.

First, the random selection strategy, *RAND*, is a benchmark to quantify the improvement of other methods.

IRT [10] is the most popular CDM in CAT, and the corresponding CAT baselines are:

- *MFI*: *Maximum Fisher Information* [12], [9] is the most popular selection strategy which measures the information of questions with the Fisher information function.
- *KLI*: *Kullback-Leibler Information* [21] is a global information heuristic that measures the informativeness with Kullback-Leibler divergence.

MIRT [11], as a multidimensional extension of IRT, shows its potential in multitrait ability estimation. In order to adapt to MIRT, the IRT-based methods were also extended. So we compare MAAT to the following baselines on MIRT:

- *D-Opt*: *D-Optimality* [23], termed in the optimization termilogy, is a multivariate extension of MFI.
- *MKLI*: *Multivariate Kullback-Leibler Information* [24] is a direct generlization of its unidimensional version, KLI.

NCDM (*Neural Cognitive Diagnosis Model*) [20] is one of the most recent deep learning CDMs. Though NCDM has shown great power, to the best of our knowledge, there is no existing methods able to work with it because of its extremely

(a) EXAM

Methods	IRT		MIRT		NCDM	
	@25	@50	@25	@50	@25	@50
RAND	0.6435	0.7076	0.7426	0.7767	0.7081	0.7566
MFI	0.7092	0.7207	-	-	-	-
KLI	0.7081	0.7257	-	-	-	-
D-Opt	-	-	0.7515	0.7710	-	-
MKLI	-	-	0.7502	0.7747	-	-
MAAT	**0.7192**	**0.7319**	**0.7600**	**0.7861**	**0.7614**	**0.7868**

(b) ASSIST

Methods	IRT		MIRT		NCDM	
	@25	@50	@25	@50	@25	@50
RAND	0.6619	0.6664	0.6734	0.6902	0.6832	0.7217
MFI	0.6659	0.6691	-	-	-	-
KLI	0.6658	0.6692	-	-	-	-
D-Opt	-	-	0.6832	0.7004	-	-
MKLI	-	-	0.6781	0.6877	-	-
MAAT	**0.6674**	**0.6703**	**0.6903**	**0.7063**	**0.7084**	**0.7334**

complex mechanism. So we compare MAAT with only RAND to show our improvement. The point is that with a comparison between the results of MAAT on different CDMs, we can validate the advantage of being model-agnostic.

D. Evaluation Metrics

We measure quality and diversity with the informativeness and coverage of the strategy (i.e., $Inf(\mathcal{S})$ and $Cov(\mathcal{S})$) respectively (Section III). In addition, we introduce a metric that has been commonly used in traditional CAT studies.

1) Informativeness Metric: Following the discussion in Section III, we measure the quality through the informativeness of the selection strategy. Specifically, for each examinee e_i in the testing data, we predict her performance on every question q_j whose ground truth has been recorded. Then we adopt the common AUC (Area Under ROC) metric:

$$Inf(\mathcal{S}) = AUC(\{M(e_i, q_j|\boldsymbol{\theta})|e_i \in E, q_j \in Q\}). \quad (17)$$

2) Coverage Metric: We measure the diversity through the coverage of the selection strategy. Since there is no universal standard, we adopt a simple form for $Cov(\mathcal{S})$, i.e., the proportion of knowledge concepts covered in the questions selected by the strategy:

$$Cov(\mathcal{S}) = \frac{1}{|K|}\sum_{k \in K} \mathbb{1}[k \in Q_T]. \quad (18)$$

3) Simulated Estimate Error Metric: Traditional CAT studies have a different evaluation process called *simulation study*, which generates imaginary data instead of using real-world data. For example, with IRT model, a group of *simulated parameters* is generated representing the ability of examinees, the difficulty of questions, etc. These parameters are used to generate imaginary records with Item Response Theory [10]. Then experimentally estimate the simulated parameters in turn with the records. To evaluate the effectiveness of the strategy, at each step in the test, we calculate the mean squared error between the estimated parameters and the simulated parameters, namely *Simulated Estimate Error* (SEE):

$$SEE(\mathcal{S}) = \frac{1}{|E|}\sum_{e_i \in E}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)^2, \quad (19)$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i^*$ are the estimated and simulated parameters related to the examinee $e_i \in E$, respectively. In traditional case studies, $\boldsymbol{\theta}_i^*$ is randomly generated as ground truth. To show that our measurement is consistent with the traditional one, we also evaluate the proposed framework with SEE metric. Because we conduct experiment on real-world data, we use the estimated parameters trained on the whole data records as $\boldsymbol{\theta}_i^*$, instead of generating them.

E. Experimental Results

1) Quality Comparison: Table III reports the comparison of quality with the informativeness metric, i.e., AUC@t (Eq. (17)). We show the results in the middle (step $t = 25$) and the end (step $t = 50$) of tests. First, MAAT is proven to be model-agnostic since it can adapt to all the CDMs. Second, we can easily see that, on both datasets, MAAT shows outperforming results with all CDMs during the test, which indicates that MAAT is fairly effective in achieving the quality goal. It is worth noting that MAAT makes no use of the details of the CDMs compared with the baseline approaches. Third, the overall results become better as the CDM becomes more complex, which is reasonable because complex CDMs do better in capturing the ability of the examinees. This observation confirms the advantage of being model-agnostic: MAAT can improve the CAT system by making it flexible to replace the CDM without redesigning the strategy.

2) Diversity Comparison: Fig. 3 illustrates the comparison of diversity with coverage metric (Eq. (18)). MAAT framework outperforms much on both datasets with all CDMs, because it has an intrinsic knowledge-level coverage goal and a performance-guaranteed optimization algorithm while other methods do not. As shown in the curve charts, the coverage of MAAT grows fairly rapidly in the early steps of tests and quickly approaches the limit of 1. This feature is very important for CAT because adaptive tests are typically short. In addition, we observe that traditional selection strategies, such as MFI and KLI, also help with the coverage goal, though they only intrinsically aim at informativeness. This observation reveals that quality and diversity are correlated instead of contradictory. Both of them can benefit the target of offering better diagnosis results for examinees.

3) Consistency Validation: Though we have evaluated the quality and diversity of MAAT, it remains important to validate the consistency of our evaluation measurement. Therefore, we conduct experiment with the Simulated Squared Error metric (Eq. (19)) additionally. The results are reported in Fig. 4. Since simulation study is only suitable for those CDMs with extremely simple and explainable parameters, we only conduct the simulation with IRT. We can see that MAAT also performs well in SEE metric.

4) Ablation Study: We conduct an ablation study on how the MAAT framework keeps a balance when quality meets
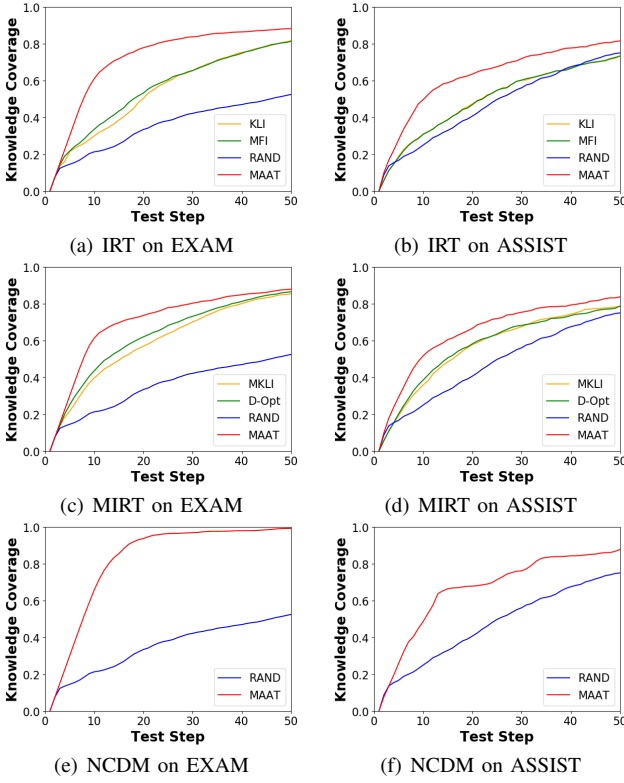
Fig. 3. Diversity Comparison with Coverage Metric

(a) IRT on EXAM
(b) IRT on ASSIST
(c) MIRT on EXAM
(d) MIRT on ASSIST
(e) NCDM on EXAM
(f) NCDM on ASSIST

Fig. 4. Simulated Estimate Error comparison

(a) EXAM
(b) ASSIST

Fig. 5. Change in quality and diversity with different K_C for ablation study

(a) Quality
(b) Diversity

diversity. The key point is the size of the candidate set, K_C, which connects Quality Module and Diversity Module. Specifically, when $K_C = 1$, only Quality Module works; when $K_C \equiv |Q_U|$, only Diversity Module works. Therefore, we observe how quality and diversity change with K_C in these boundary conditions. Due to limited pages, we show the results with NCDM on EXAM as an example in Fig. 5. The change in diversity is straightforward: the larger K_C, the faster the coverage metric increases. Specially, when only Diversity Module works ($K_C \equiv |Q_U|$), we achieve the performance guarantee discussed in Section IV-C. Moreover, we observe that the diversity quite approaches the theoretical limit when K_C is a small value (i.e., $K_C = 10$). Note again that when only Quality Module works (i.e., $K_C = 1$), there is still improvement on diversity compared with RAND benchmark, because the two goals are correlated. The change in quality is slightly more interesting. The case $K_C = 1$ does not always perform the best because Diversity Module and Importance Module can also help with quality by taking coverage and importance into consideration. To sum up, a relatively small K_C keeps the best balance of quality and diversity.

5) Case Study: We present the first 10 steps of a typical examinee in EXAM for case study (Table IV). For better illustration, we only compare MAAT with the best baselines in the previous experiment, i.e., D-Opt and MKLI on MIRT. For each methods, we show the associated knowledge concepts in the first column and the informativeness metric, i.e. AUC@t (Eq. (17)), at the corresponding step in the second column. We abbreviate the knowledge concepts, such as "Linear Equation
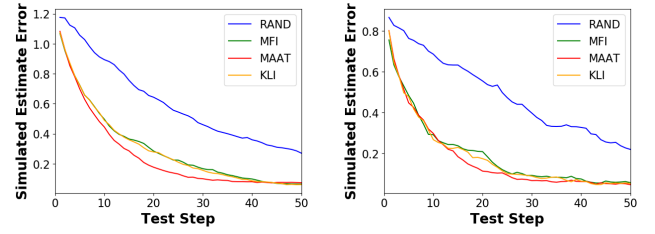
in One Variable" to "Equation". We can clearly see that, our MAAT framework can select diverse questions while keeping high quality as well. Specifically, with a slightly outperforming AUC, MAAT covered 9 knowledge concepts with the first 10 selected questions, while D-Opt and MKLI covered only 5 and 6 respectively. Moreover, the two baselines tended to select more questions about "Function", while MAAT did not. Therefore, our MAAT framework makes better selections with a good balance of quality and diversity.

VI. CONCLUSIONS

In this paper, we studied a novel model-agnostic CAT problem. We proposed a novel *M*odel-*A*gnostic *A*daptive *T*esting framework (MAAT) for the solution with addressing the problem of selecting both high-quality and diverse questions in the testing procedure. In MAAT, we designed three sophisticated modules that worked cooperatively and iteratively. At each selection step, Quality Module firstly selected a small candidate set of the most informative questions with EMC score function. Diversity Module then selected one question from the candidates maximizing the knowledge coverage via IWKC, where the importance weights were pre-computed in Importance Module. Moreover, we proved that our problem was NP-hard, and provided an efficient and effective solution by the submodular property. Extensive experiments demonstrated that MAAT was flexible for any CDMs and could generate both high-quality and diverse questions in CAT. We hope this work can lead to more studies in the future.

VII. ACKNOWLEDGEMENT

TABLE IV
RESULTS ON A TYPICAL EXAMINEE FOR CASE STUDY

MAAT		D-Opt		MKLI	
Concept	Inf	Concept	Inf	Concept	Inf
Function	0.6666	Function	0.6652	Triangle	0.6645
Set	0.6710	Equation	0.6686	Algebra	0.6689
Equation	0.6763	Equation	0.6717	Equation	0.6732
Triangle	0.6841	Triangle	0.6756	Function	0.6774
Algebra	0.6905	Geometry	0.6801	Algebra	0.6810
Triangle	0.6961	Function	0.6857	Function	0.6843
Coordinates	0.7022	Geometry	0.6914	Function	0.6887
Geometry	0.7087	Triangle	0.6956	Triangle	0.6929
Real Number	0.7136	Algebra	0.6963	Inequality	0.7001
Equation	0.7188	Function	0.6998	Geometry	0.7057

REFERENCES

[1] S. Minn, Y. Yu, M. C. Desmarais, F. Zhu, and J.-J. Vie, "Deep knowledge tracing and dynamic student classification for knowledge tracing," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1182–1187.

[2] M. Hang, I. Pytlarz, and J. Neville, "Exploring student check-in behavior for improved point-of-interest prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 321–330.

[3] D. Magis, D. Yan, and A. A. Von Davier, *Computerized adaptive and multistage testing with R: Using packages catr and mstr*. Springer, 2017.

[4] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy, *Computerized adaptive testing: A primer*. Routledge, 2000.

[5] L. M. Rudner, "Implementing the graduate management admission test computerized adaptive test," in *Elements of adaptive testing*. Springer, 2009, pp. 151–165.

[6] C. N. Mills and M. Steffen, "The gre computer adaptive test: Operational issues," in *Computerized adaptive testing: Theory and practice*. Springer, 2000, pp. 75–99.

[7] W. J. van der Linden and C. A. Glas, *Elements of adaptive testing*. Springer, 2010.

[8] J.-J. Vie, F. Popineau, É. Bruillard, and Y. Bourda, "A review of recent advances in adaptive assessment," in *Learning analytics: fundaments, applications, and trends*. Springer, 2017, pp. 113–142.

[9] H.-H. Chang, "Psychometrics behind computerized adaptive testing," *Psychometrika*, vol. 80, no. 1, pp. 1–20, 2015.

[10] S. E. Embretson and S. P. Reise, *Item response theory*. Psychology Press, 2013.

[11] T. A. Ackerman, M. J. Gierl, and C. M. Walker, "Using multidimensional item response theory to evaluate educational and psychological tests," *Educational Measurement: Issues and Practice*, vol. 22, no. 3, pp. 37–51, 2003.

[12] F. M. Lord, *Applications of item response theory to practical testing problems*. Routledge, 1980.

[13] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1183–1192.

[14] J.-J. Cai, J. Tang, Q.-G. Chen, Y. Hu, X. Wang, and S.-J. Huang, "Multi-view active learning for video recommendation," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 2053–2059.

[15] D. Tu, S. Wang, Y. Cai, J. Douglas, and H.-H. Chang, "Cognitive diagnostic models with attribute hierarchies: Model estimation with a restricted q-matrix design," *Applied psychological measurement*, vol. 43, no. 4, pp. 255–271, 2019.

[16] C.-J. Lin and H.-H. Chang, "Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing," *Educational and psychological measurement*, vol. 79, no. 2, pp. 335–357, 2019.

[17] J. d. l. Torre and N. Minchen, "Cognitively diagnostic assessments and the cognitive diagnosis model framework," *Educational Psychology*, vol. 20, no. 2, pp. 89–97, 2014.

[18] Z. Huang, Q. Liu, Y. Chen, L. Wu, K. Xiao, E. Chen, H. Ma, and G. Hu, "Learning or forgetting? a dynamic approach for tracking the

knowledge proficiency of students," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 2, pp. 1–33, 2020.

[19] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Advances in neural information processing systems*, 2015, pp. 505–513.

[20] F. Wang, L. Qi, E. Chen, Z. Huang, C. Yuying, Y. Yin, and S. Wang, "Neural cognitive diagnosis for intelligent education systems," in *Thirty-Forth AAAI Conference on Artificial Intelligence*, 2020.

[21] H.-H. Chang and Z. Ying, "A global information approach to computerized adaptive testing," *Applied Psychological Measurement*, vol. 20, no. 3, pp. 213–229, 1996.

[22] C. Wang and H.-H. Chang, "Item selection in multidimensional computerized adaptive testing—gaining information from different angles," *Psychometrika*, vol. 76, no. 3, pp. 363–384, 2011.

[23] G. Hooker, M. Finkelman, and A. Schwartzman, "Paradoxical results in multidimensional item response theory," *Psychometrika*, vol. 74, no. 3, pp. 419–442, 2009.

[24] L. M. Rudner, "An examination of decision-theory adaptive testing procedures," in *annual meeting of the American Educational Research Association*, 2002.

[25] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in *Advances in neural information processing systems*, 2010, pp. 892–900.

[26] P. Bachman, A. Sordoni, and A. Trischler, "Learning algorithms for active learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 301–310.

[27] D. Ting and E. Brochu, "Optimal subsampling with influence functions," in *Advances in Neural Information Processing Systems*, 2018, pp. 3650–3659.

[28] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 510–520.

[29] R. Sipos, A. Swaminathan, P. Shivaswamy, and T. Joachims, "Temporal corpus summarization using submodular word coverage," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 754–763.

[30] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and J. X. Yu, "Influence maximization over large-scale social networks: A bounded linear approach," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 171–180.

[31] L. Wu, Q. Liu, E. Chen, N. J. Yuan, G. Guo, and X. Xie, "Relevance meets coverage: A unified framework to generate diversified recommendations," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, pp. 1–30, 2016.

[32] M. Hammar, R. Karlsson, and B. J. Nilsson, "Using maximum coverage to optimize recommendation systems in e-commerce," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 265–272.

[33] Z. Huang, Q. Liu, C. Zhai, Y. Yin, E. Chen, W. Gao, and G. Hu, "Exploring multi-objective exercise recommendations in online education systems," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 1261–1270.

[34] H. Wang, E. Chen, Q. Liu, T. Xu, D. Du, W. Su, and X. Zhang, "A united approach to learning sparse attributed network embedding," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 557–566.

[35] L. Zhang, X. Wu, H. Zhao, F. Cheng, and Q. Liu, "Personalized recommendation in p2p lending based on risk-return management: A multi-objective perspective," *IEEE Transactions on Big Data*, 2020.

[36] Q. Liu, Y. Ge, Z. Li, E. Chen, and H. Xiong, "Personalized travel package recommendation," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 407–416.

[37] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[38] O. Barkan and N. Koenigstein, "Item2vec: neural item embedding for collaborative filtering," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.