Multi-Objective Optimization of Item Selection in Computerized Adaptive Testing

Dena F. Mujtaba and Nihar R. Mahapatra
Department of Electrical and Computer Engineering, Michigan State University
East Lansing, Michigan, USA
{mujtabad,nrm}@egr.msu.edu

ABSTRACT

Computerized-adaptive testing (CAT) is a form of assessment in which items/questions are administered based upon a test taker's ability (i.e., their estimated proficiency, such as a knowledge, skill, or personality characteristic). CAT is regularly used in psychological studies, medical exams, and standardized testing to reduce test length and improve measurement accuracy and precision. A key challenge in CAT is item selection. Past algorithms have been designed based on criteria such as item difficulty and maximum Fisher information. However, these only consider a fixed-length test, which may result in it being longer or less precise. To address this problem, we formulate a new multi-objective optimization problem to model the trade-off between test length and precision. A binary population-based genetic algorithm, NSGA-II, is used to obtain the set of Pareto-optimal solutions by maximizing precision and minimizing the number of questions. We evaluate our approach with a simulated study using four standard personality assessments. We also investigate the influence of test response types (e.g., binary versus categorical response) and number of variables (i.e., the number of possible items) on performance. The results obtained show multi-objective optimization can be used in CAT to minimize overall test length and improve measurement precision and overall accuracy.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Psychology**; **Multi-criterion optimization and decision-making**.

KEYWORDS

computerized adaptive testing, computational psychometrics, item response theory, multi-objective optimization, NSGA-II

ACM Reference Format:
Dena F. Mujtaba and Nihar R. Mahapatra. 2021. Multi-Objective Optimization of Item Selection in Computerized Adaptive Testing. In *2021 Genetic and Evolutionary Computation Conference (GECCO '21), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3449639.3459334

GECCO '21, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8350-9/21/07…$15.00
https://doi.org/10.1145/3449639.3459334

1 INTRODUCTION

1.1 Motivation and Background

Computerized-adaptive testing (CAT) is a form of electronic assessment in which items/questions are administered based upon a test taker's ability. The responses to these questions are used to estimate proficiency with respect to a *latent trait* dimension, denoted by θ, that is not directly observable (such as general intelligence, knowledge, skill, ability, or personality characteristic) [21]. CAT seeks to address many of the challenges faced by survey researchers administering assessments [21]. CAT is able to provide a higher level of precision while reducing test length in comparison to static assessments (in which all test questions are included to achieve the highest level of precision) and static-reduced assessments (in which only a pre-selected subset of questions are included). This is because CAT efficiently customizes each assessment's set of questions to a test taker's ability, excluding excessively difficult questions that may force the test taker to guess and so provide little additional information about their ability [21, 37]. Moreover, reducing test length often reduces cost for survey researchers, since study participants take less time to complete the assessment. In addition, finding participants will be easier for study administrators since past studies have shown that individuals are more likely to complete tests that are shorter in length or less time-consuming [15]. Overall, CAT is a crucial tool for survey researchers because assessments are a widely-used research method in psychological studies, medical exams, and other areas, and are currently used in many standardized tests such as the Graduate Record Examination and the U.S. Department of Defense's Armed Services Vocational Aptitude Battery [21]. CAT can greatly improve exams for students.

The structure of a standard CAT is shown in Figure 1. First, an initial set of items (i.e., questions) is established to measure one or more latent traits [21, 37]. Then, an item selection algorithm is used to select an item to serve the test taker in each iteration. If the test taker answers the item correctly, the estimated ability is seemingly higher, whereas if they answer incorrectly, the estimated ability is lower. The ability is estimated using different psychometric models, and more recently machine learning and deep learning models have been used. The new ability is used by the selection algorithm to choose the next best item to serve the test taker, until a termination criterion is met (e.g., the maximum number of items have been served or the estimated error is lower than a threshold).

An important challenge in CAT is item selection. Past algorithms have been designed based on criteria such as item difficulty, maximum Fisher information, Kullback-Leibler information, and others. However, these only consider a fixed-length test, which may result in a longer or less precise test [16]. Furthermore, selecting only the

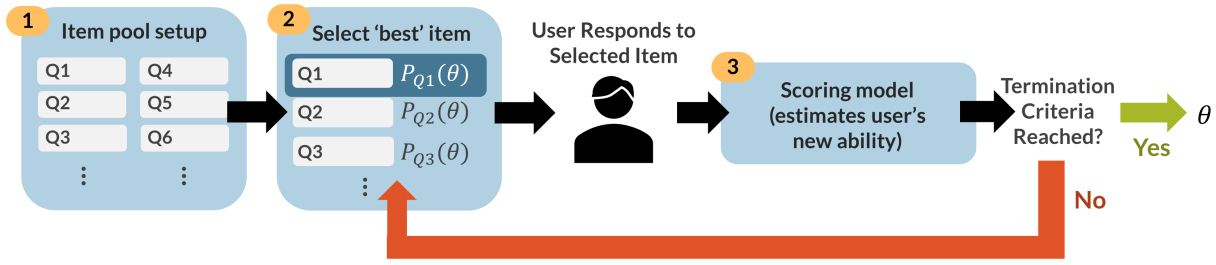

Figure 1: Overview of steps in adaptive testing. Here, θ refers to the ability of the test taker, i is an item, and $P_i(\theta)$ is the probability of a correct/contributing response, a commonly used factor in item selection.

single-most optimum next item to serve in CAT can be resource-intensive. Item selection also heavily relies on *item response theory* (IRT), which models the probability of a correct response to an item as a function of person parameters (such as math ability) and item parameters (such as difficulty) [38] to choose the next best item to serve the test taker [8, 38]. IRT and other design parameters (assessment length, selection method, etc.) create a higher barrier to entry for survey researchers wanting to use CAT due to the long implementation time and need for psychometricians familiar with statistical models to validate the final CAT parameters [7, 17].

1.2 Key Contributions

To address these problems, we formulate a new multi-objective optimization problem to model the trade-off between test length and precision. The optimization and evolutionary-algorithms (EA) community has looked at problems in sociology, economics, and other areas of the social sciences. However, there has not been any work on EAs for CAT or psychometrics. Furthermore, past works in psychometrics and CAT have studied item selection approaches, but not with EAs or multi-objective optimization, and instead use a single-objective and greedy strategy for selection in each iteration. To our knowledge, this is the first work that applies EAs for CAT and item selection.

The main contributions of this paper relative to previous work are fourfold: (1) formulation of a new multi-objective optimization problem to assist survey researchers by supporting decision-making in CAT design, allowing for trade-off between test length and precision, (2) use of multi-objective evolutionary algorithms (MOEA) to obtain the set of Pareto-optimal solutions by maximizing precision and minimizing the number of questions, (3) an analysis of the effect of different assessment characteristics on the evolutionary solution approach, and (4) evaluation of the new multi-objective optimization approach using simulated CAT studies with four standardized personality assessments, different IRT models, and varying number of questions. By modeling both test length and precision at each iteration, a more resource-efficient approach and accurate measurement can be achieved for the overall test.

The remainder of this paper is organized as follows. In Section 2, we provide a literature review of item selection approaches, and use of artificial intelligence (AI) in all stages of CAT. Next, Section 3 presents our problem formulation and Section 4 details the methods and data used in our approach. Then, Section 5 describes the experiments, evaluation, and results obtained from our approach. Last, Section 6 provides conclusions and highlights future work.

2 RELATED WORK

CAT research, starting in the 1970s, has expanded over the past few decades [21]. However, only recently has AI, such as natural language processing and machine learning, been included and studied. In this section, we provide a literature review of item selection methods, and aspects of CAT where AI has been introduced.

2.1 Item Selection Methods

Many item selection criteria have been proposed in the past for CAT design [16]. Selection algorithms heavily rely on IRT models and the estimated ability of the test taker [21, 34, 38]. These models take the form [20]:

$$\mathbb{P}(U = u|\theta) = f(\theta, \eta, u), \tag{1}$$

where $\mathbb{P}$ is the conditional probability that a test taker characterized by a vector of latent trait parameters θ will respond to the test item with value u (e.g., $u \in \{0, 1\}$ or $\{u \in \mathbb{Z} : 1 \leq u \leq 5\}$). $\mathbb{P}$ is defined by a function f that relates θ and a vector η of parameters characterizing the item [20]. The test taker's probable response is used by item selection algorithms to select the next item to serve the test taker. Two commonly used approaches are maximum information and Bayesian item selection, or the minimum expected posterior variance [16, 21]. Other selection criteria include: difficulty matching in which items are selected based on the distance between θ and item pool difficulty parameters, a-stratification in which items are grouped based on their parameters prior to selection, and Kullback-Leibler information in which a moving average for any θ is used. There are also many other contributing factors in item selection, such as content balancing, the criteria used to select items, and item exposure control. A comprehensive review of traditional item selection algorithms and strategies is further described in the survey by Han *et. al* [16].

2.2 AI in Computerized Adaptive Testing

AI has been incorporated to provide more efficient and accurate results in each stage of CAT, and can be categorized into three types: automatic item generation, item selection, and scoring algorithms. First, *automatic item generation* (AIG) seeks to generate question prompts similar to those found in other assessments, and

has relied on natural language processing to provide state-of-the-art results [4, 5, 25, 30, 32, 36, 37]. However, this is still an emerging area and there are many challenges faced by researchers such as generating text with the correct syntax and semantics, capturing domain-specific questions for each test, generating according to the reading comprehension and ability of the test taker, and IRT modeling for AI-generated questions [10]. Next, in item selection, recommendation methods such as collaborative-filtering [39] and contextual-bandits [19] have been used. Furthermore, the approach by Li *et al.* uses deep Q-learning to estimate test taker latent ability and recommend exercises for online learning and coursework [20]. Last, many scoring algorithms have used deep learning and natural language processing to learn from large datasets and improve accuracy [4, 9, 18, 29, 33, 40, 41]. The model DIRT (deep item response theory) estimates student latent trait ability using a proficiency vector of knowledge areas and a deep neural network to predict the latent trait, discrimination factor, and difficulty for IRT [9]. Another model, Deep-IRT, also uses deep learning to predict student latent ability [41]. Most work in ability estimation has focused on student learning, which is also known as *knowledge tracing*. A review of knowledge tracing and ability estimation models can be found in the survey by Khajah *et al.* [18].

AI can be a powerful method for CAT to produce flexible tests. However, this field is relatively new and has not been adopted in psychological practice because of its high barrier to entry for design and implementation [7, 17]. Our approach seeks to assist survey researchers in CAT design and reduce implementation costs by providing a new multi-objective optimization strategy with MOEAs to improve item selection, the core component of CAT. Our approach uncovers the Pareto-efficient front to visualize the trade-off between test length and precision.

3 PROBLEM FORMULATION

We formulate the optimization problem with past IRT models to capture the ability of each individual test taker. We minimize the length of the test (i.e., the number of questions) and maximize the precision, or minimize the standard error of measurement, as further described next.

Number of questions: Each question in the full assessment, or item bank, is denoted by binary variable x, where $x = 1$ if a question is present, and $x = 0$ is a question is not present for the test taker. The total number of questions in the item bank, or variables in the optimization problem, is determined by the four assessment datasets used in our simulated CAT study. To evaluate different scenarios, we use tests with varying lengths and characteristics, as described in Section 4.1.

For each solution, the length of the test in one CAT iteration is given as:

$$f_1(x) = \sum_{i=1}^{N} x_i, \tag{2}$$

where N is the total number of questions in the item bank.

Standard error of measurement: The precision, or *standard error of measurement* (SEM), is the confidence in an estimate from a test. As test length decreases, the SEM generally increases. We use SEM as the second objective over other item selection criteria because of its wide use as the termination criteria for CAT-based studies that have also been validated [21, 23]. SEM is given by:

$$f_2(\theta, x) = \sqrt{\frac{1}{\sum_{i=1}^{N} I_i(\theta, b_i, a_i) x_i}}, \tag{3}$$

where the function I represents the information for an item i for a test taker with ability θ in the test, b_i is a question difficulty parameter in which $b_i = 1$ indicates high difficulty and $b_i = 0$ indicates low difficulty, and a_i is a discrimination parameter in which a higher value means the question can easily differentiate abilities of test takers. Item information is given by:

$$I_i(\theta, b_i, a_i) = a_i^2 \mathbb{P}_i(\theta, b_i, a_i) Q_i(\theta, b_i, a_i), \tag{4}$$

where $Q_i(\theta, b_i, a_i) = 1 - \mathbb{P}_i(\theta, b_i, a_i)$ and $\mathbb{P}_i(\theta, b_i, a_i)$ is an IRT model defining the probability of the test taker answering question i correctly [16, 35]. This model depends on the test and the type of question (e.g., multiple choice or Likert scale), as further described next.

Item response theory models: There are two types of items among the datasets used for evaluation: *dichotomous* items, in which a binary response is given, such as True/False, and *polytomous* items in which a categorical response is given, such as a one to five scale. For dichotomous items, the first model was described in 1960, called the Rasch model, or the 1PL (one-parameter logistic) model. This defines the simplest model to summarize a person's ability and is given by [20, 26, 38, 38]:

$$\mathbb{P}(U_{i,j} = 1 | \theta, b_i) = \frac{1}{1 + e^{-(\theta - b_i)}}, \tag{5}$$

where $U_{i,j}$ is the response of the i^{th} item by the j^{th} test taker. This model has also been extended to include additional parameters. The 2PL (2-parameter-logistic) model [31] for binary response items is given by:

$$\mathbb{P}_i(U = 1 | \theta, b_i, a_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}} \tag{6}$$

where U is the response of the test taker and $U = 1$ indicates a correct response. Other models like the 3PL and 4PL model have also been defined. However, depending on the assessment and the dataset size, a model will need to be chosen accordingly (e.g., for a smaller dataset, the 3PL model would not be used because it requires more data to accurately infer ability with more parameters) [38]. Moreover, the 2PL model has been used extensively in past studies and validated over time [24, 26, 34]. Therefore, we use the 2PL model for the dichotomous item datasets. An example of the resulting SEM, item information, and the 2PL model for a sample question from the datasets used is given in Figure 2.

In contrast, polytomous item models represent items with categorical responses, such as a Likert question with a scale of 1 to 5 [24]. The advantage of a categorical response over binary is more precise trait estimates that can be obtained from the test taker [24]. Many models were developed to evaluate responses such as the graded response model, nominal response model, and the partial credit model [24, 34, 42]. We use the *graded response model* (GRM) [24, 42], which builds upon the dichotomous item model. GRM expands upon the 2PL model by considering the probability $\mathbb{P}_{i_z}(\theta)$

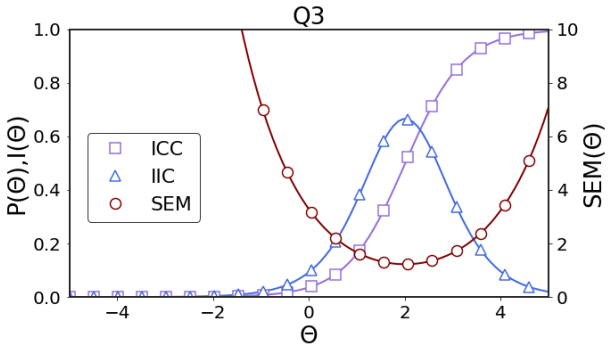

Figure 2: Item characteristic curve (ICC) showing the probability of a correct response, item information curve (IIC), and the SEM for dichotomous items [27].

of the test taker's response for item i beginning above or below θ for a category z as given by [24, 42]:

$$\mathbb{P}_{i_z}(\theta) = \frac{e^{a_i(\theta - t_{i_z})}}{1 + e^{a_i(\theta - t_{i_z})}}, \tag{7}$$

where t_{i_z} is the threshold parameter for the latent trait level where the probability of the test taker responding at or above category z is 50%. An example of the GRM model from a sample question from the datasets used is given in Figure 3.

All models have been used extensively in other work and their reliability and validity studied with both simulated psychological and real-world studies [24, 26, 34]. The parameters a and b for every item will be calibrated using data from other studies and their assessments, further described in Section 4.1.

Ability estimate: Initially, θ is set to 0, and is estimated after each question the user answers. We use the expected *a posteriori* (EAP) estimate to calculate the test taker's position on the latent scale. EAP is the most commonly used estimate in adaptive testing [22, 23]. This is given by:

$$\hat{\theta}^{(EAP)} \equiv E(\theta|U_{k-1}) = \frac{\int \theta\pi(\theta)L(\theta|U_{k-1})d\theta}{\int \pi(\theta)L(\theta|U_{k-1})d\theta}, \tag{8}$$

where a prior distribution for θ is given by $\pi(\theta) \sim N(\mu_\theta, \frac{1}{\tau_\theta})$, and τ_θ denotes the precision of the distribution and both τ_θ and μ_θ are defined based on the assessment datasets used [22, 23]. L is the likelihood function for the questions answered $k-1$. This is given by:

$$L(\theta|U_{k-1} = u) = \prod_{i=1}^{k-1} \mathbb{P}_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}, \tag{9}$$

for dichotomous items, where $\mathbb{P}_i(\theta)$ and $Q_i(\theta)$ refer to the IRT models described earlier. For polytomous items, the likelihood function is given by:

$$L(\theta|U_{k-1} = u) = \prod_{i=1}^{k-1} \prod_{t=1}^{C_i} \mathbb{P}_{it}(\theta)^{I(u_i = t)}, \tag{10}$$

where C are all possible categories for a response and $I(\cdot)$ is an indicator function [23].

An example plot of the estimated ability over each iteration of a question is shown in Figure 4 for a dichotomous item type, and Figure 5 for a polytomous item type. The estimated ability θ is shown on the y-axis, and the question given to the test taker is shown on the x-axis. The difficulty parameter for each question is also shown in red. A green background indicates the question was answered correctly, and red indicates it was answered incorrectly. We observe that when the test taker answers the question correctly, the estimated ability increases, and when answering incorrectly the ability decreases. Towards the end of the assessment it converges to the individual's real ability.

Optimization problem: At each iteration of the CAT, the overall optimization problem solved to obtain the Pareto-optimal set of solutions is given by:

$$\min_{i=0,\ldots,N} [f_1(x_i), f_2(x_i, \theta)], \tag{11}$$

subject to the constraint $-\sum_{i=1}^{N} x_i + 2 \leq 0$, which states that there must be at least two questions given to the user in the test, allowing for a more accurate estimate of a person's ability (i.e., there cannot be a one question survey).

The number of variables N is equal to the number of questions in the full assessment, and decreases as questions are given to the test taker. At each step of the CAT, a solution defining the items to serve the test taker, is selected from the Pareto-optimal set. Once items are served, they are removed from the item bank, meaning N will decrease each time the optimization problem is solved in the CAT. Therefore, we add another constraint, or a stopping condition for the test, that the CAT will terminate once the item bank is exhausted, or when the standard error of measurement indicates 95% confidence or higher.

4 MATERIALS AND METHODOLOGY

CAT results depend upon several characteristics of the assessment, such as the length, number of traits being measured, distribution of abilities, and other factors. In this section, we explore the datasets used to calibrate the IRT models and evaluate the MOEA used to solve the optimization problem, and the decision-making strategy.

4.1 Data & Model Calibration

To calibrate the IRT models and compute SEM, we use data from previous psychological studies where individuals have already completed a full version of an assessment. We use four standardized assessments as training data: the Narcissistic Personality Inventory (NPI) [27], the Machiavellianism Personality Test (MACH) [11], the

Table 1: Datasets used and characteristics.

Dataset	N	IRT model	Response categories	Records
NPI	40	2PL	2	11,200
MACH	20	GRM	5	58,700
EQSQ	60	GRM	4	13,200
POL	65	2PL	2	860

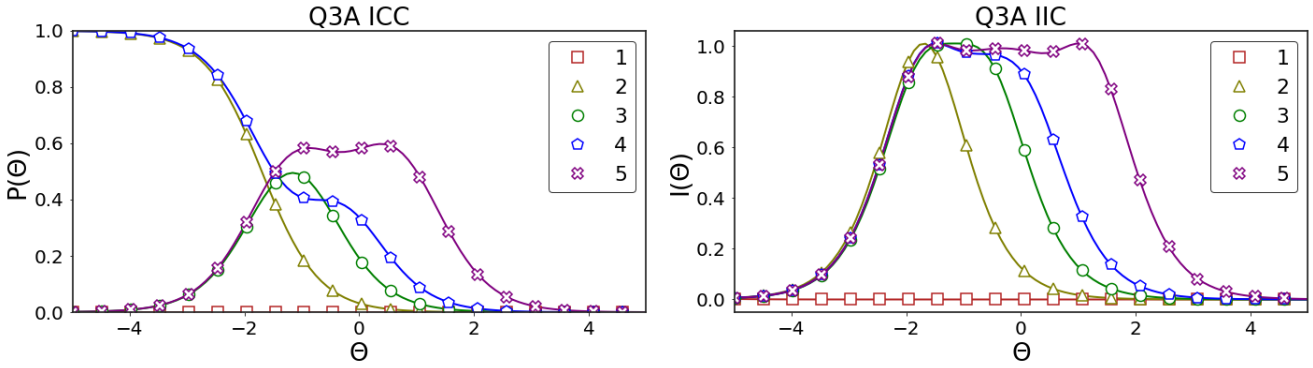

Figure 3: Item characteristic curves for each type of response category in polytomous items [11].

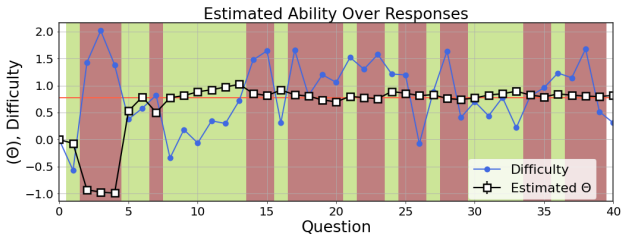

Figure 4: Estimated ability and question difficulty over an entire dichotomous 40-question test. The estimated ability increases when the test taker answers correctly (shown in green), and decreases when answering incorrectly (shown in red). Over many responses, the ability converges to the actual value (indicated by the red line).

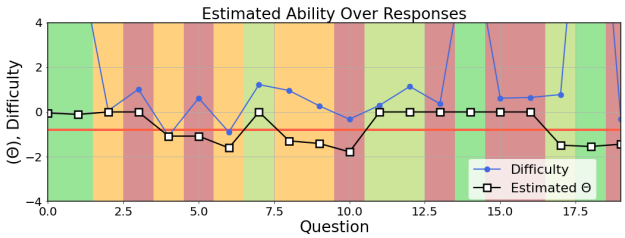

Figure 5: Estimated ability for a polytomous test, where responses range from 1 ("strongly disagree") to 5 ("strongly agree"). The color scale from red indicates 1 to green indicating 5. Yellow indicates a neutral response, 3.

Empathy and Systemizing Quotient [2], and a political psychology/sophistication assessment (POL) [22]. All datasets used are from the Open-Source Psychometrics Project, found in [1], and in [22] for POL. The Open-Source Psychometrics Project collects data from online surveys, and POL data was collected through Mechanical Turk; both have been shown to provide similarly valid and reliable results to in-person studies [3].

Example questions from the POL and MACH tests are shown in Figure 6. The *R* library *ltm* was used to calibrate the model parameters, such as difficulty and discrimination, as well as find the

Figure 6: Example of a question from the POL dataset and one from the MACH personality dataset.

distribution of scores in the dataset [28]. We use 80% of the dataset for calibration and 20% for the remaining results and evaluation. The IRT parameters, score distribution, and questions in each assessment are further described in the Supplementary Documents. Dataset characteristics are given in Table 1.

4.2 Multi-Objective Evolutionary Algorithm

MOEAs are used to find the set of Pareto-optimum solutions for multi-objective optimization problems, and can model the trade-off between different objectives [12]. MOEAs are also ideal for finding a diverse solution set to represent the entire Pareto-optimal front. One of the most popular implementations is the fast Elitist Non-Dominated Sorting Genetic Algorithm II (NSGA-II) [13]. NSGA-II is a fast approach with a diversity preserving mechanism to find solutions closest to the Pareto-optimal front [13]. After population initialization and selection, crossover, and mutation, a non-dominated sorting operation is performed, which groups the population into fronts. Then, a crowding distance operation is performed, which assigns a crowding distance value to population individuals in each front as the perimeter of the cuboid, or objective space around the individual unoccupied by any other solution in the population. The next generation population is formed by selecting individuals from the first fronts and solutions with a higher crowding distance value. NSGA-II is further described in [12, 13].

We use NSGA-II and the implementation in the Python library *pymoo* [6] to find the Pareto-optimal set of solutions. Each solution is represented as a binary vector $X = \{x_1, x_2, \ldots, x_i, \ldots, x_N\}$ where

$x_i \in \{0, 1\}$ and N is the total number of questions in the item bank. A two-point crossover operator and a bit flip mutation operator is used [14]. We use a binary representation and crossover and mutation operators because of its computationally cost-efficient implementation. Duplicate solutions are also removed from the population in each generation. Each iteration of the CAT also removes any questions from the item bank that were given to the test taker, therefore reducing N over time; this does not affect the number of solutions in each generation, which relies on the pre-defined population size parameter. The mutation rate, crossover rate, and other parameters for the MOEA are given in Section 5.

4.3 Decision-Making

Next, we use the pseudo-weight method to select a solution of questions from the Pareto-optimal set [12]. Pseudo-weight vectors are assigned to each solution to capture the context of the solution in multi-objective optimization. The pseudo-weight vector implementation used is from the Python library pymoo [6]. The pseudo weight for the i^{th} objective is given by the following:

$$w_i = \frac{(f_i^{max} - f_i(x))/(f_i^{max} - f_i^{min})}{\sum_{m=1}^{M}(f_m^{max} - f_m(x))/(f_m^{max} - f_m^{min})}, \tag{12}$$

where M is all solutions in the Pareto-optimal set and f_i is the i^{th} objective function [6]. The advantage this method provides is the simplicity in selecting a solution (e.g., giving 80% weightage for f_1 and 20% for f_2, meaning the number of questions is given higher priority than precision, and will therefore result in a shorter test).

This is beneficial for a real study, where the survey researcher running the study would decide the length of the test, or the trade-off between the SEM and number of questions. Moreover, a pseudo-weight method supports the survey researcher to select a solution for each iteration of the adaptive test, with the context of the entire Pareto-optimal front. A selection alternative could also allow the test taker to decide the length of the test they take, while understanding the trade-off of precision (e.g., in hiring tests). However, to evaluate the results in this project, we observe three different solutions with different weights: $W_1 = (0.25, 0.75)$, $W_2 = (0.5, 0.5)$, and $W_3 = (0.75, 0.25)$. This will allow us to analyze the trade-off between precision and test length.

5 RESULTS AND DISCUSSION

Our results consist of three parts. First, we calibrate IRT models for computing SEM and estimated θ; final IRT parameters are given in the Supplementary Documents to reproduce results. This is used to obtain the Pareto-optimal fronts, covered in Section 5.1, and evaluated with single-objective optimization for each objective. Then, we use a pseudo-weight method with three different weight combinations to find different solutions on the front, and discuss the decision-making strategy when using the adaptive test in a real study in Section 5.2. Last, the final item selection algorithm is evaluated in Section 5.3.

For each optimization run, we set population size to twice the size of the assessment item bank and run for 200 generations. Therefore, for an assessment with 60 questions, such as EQSQ, a total of 1,200 solutions are simulated and evaluated. The GA mutation rate is set at 5%, crossover at 100%, and starting θ for IRT is set to 0. These parameters were selected based upon convergence of the GA with different parameters and experiments.

5.1 Non-Dominated Set of Solutions

NSGA-II is used to obtain Pareto-optimal solutions. This is shown in Figure 7 for all traits in each dataset. We are minimizing both the SEM, given by F2, and the number of questions, given by F1. For each front, we perform 10 independent runs and combine the non-dominated solutions from all runs to obtain the final front.

We verify the results obtained by optimizing/minimizing each objective at a time. The results are shown as the green and diamond points in Figure 7. For SEM, the minimum value found was 0, where all the questions in the assessment were included. For the number of questions, the minimum number was 2 (due to the constraint set). However, in the Pareto front we notice the solutions containing few questions are very sparse. This is due to the question dataset and parameters. Questions have different IRT models, and will therefore lead to different SEM. Looking at the ability estimate plot with question difficulty in Figure 4, we notice a select few questions with a large difficulty parameter, meaning the test taker will likely respond incorrectly or guess. Since these questions were included in the objective space, they had a much higher SEM, setting it apart from other solutions as a dominated solution.

5.2 Preference-Based Decision-Making

Next, we use the pseudo-weight method to select different solutions from the Pareto-optimal set. The selected solutions are shown in red in Figure 7 on the Pareto-fronts, and detailed in Table 2. We observe that the Pareto-front for EQSQ and NPI reflect a steep increase in questions when using different pseudo-weights. However, a survey researcher may use this to their advantage, by understanding the trade-off between test length and SEM, fewer questions can be included while still achieving a low SEM. For the pseudo-weights W_1, a larger assessment was selected, such as one with 36 questions for both EQSQ and 14 for MACH. However, the opposite was for pseudo-weights W_3, where only 2 questions were given to the test taker in all datasets. Moreover, we observe that the SEM for dataset MACH and EQSQ had much lower SEM values than the POL or NPI datasets. This is due to the question parameters and the polytomous IRT model, since MACH and EQSQ had categorical responses. We use the selected solutions for evaluation of each dataset, as further described next.

Table 2: Final selected solutions from the Pareto-optimal sets with different weight vectors. $f_1(x)$ is the number of questions and $f_2(x)$ is the SEM.

	Selected Solution					
	$W_1 : (0.25, 0.75)$		$W_2 : (0.5, 0.5)$		$W_3 : (0.75, 0.25)$	
Trait	$f_1(x)$	$f_2(x)$	$f_1(x)$	$f_2(x)$	$f_1(x)$	$f_2(x)$
NPI	27	0.0019	11	0.0132	2	0.0441
MACH	14	0.0028	6	0.0027	2	0.0085
EQSQ-E	36	0.0002	11	0.0041	2	0.0165
EQSQ-S	36	0.0002	11	0.0042	2	0.0165
POL	38	0.0009	12	0.0113	2	0.0421

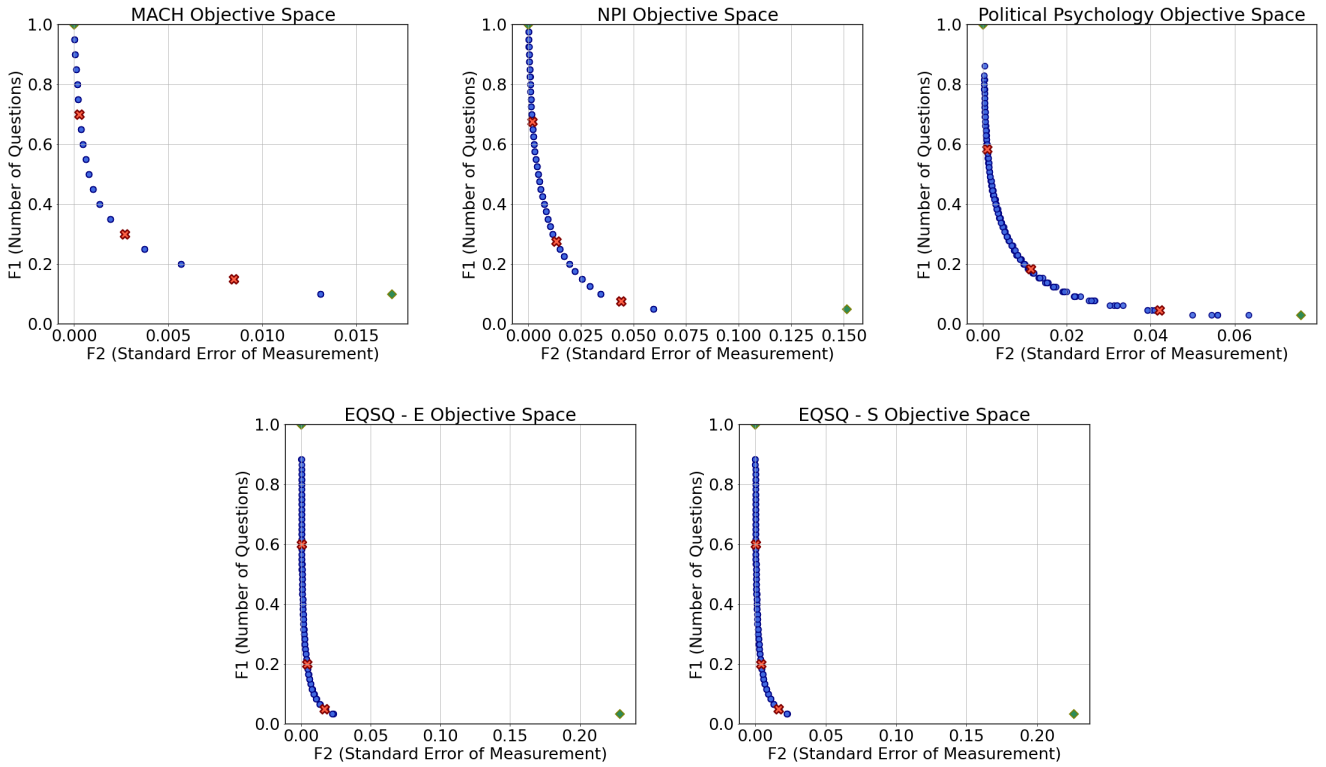

Figure 7: The final Pareto-optimal fronts obtained. We run ten independent simulations with different seeds and select the non-dominated solutions. Both objectives have also been normalized. We note that the EQSQ has two separate Pareto-optimal fronts because the assessment is multidimensional and measures two different traits, each with 60 questions.

5.3 Simulated Evaluation

To evaluate the adaptive version of the assessment, selected solutions, overall optimization problem, and compare performance across datasets and experiments, we use the root mean squared error (RMSE) metric, which compares the estimated responses from the full assessment to the adaptive version. This metric is a standard used in CAT to compare the adaptive assessment to the static assessment [23]. RMSE is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2}{n}} \tag{13}$$

where for n is the total number of predictions and θ_i is the observed ability and $\hat{\theta}_i$ is the estimated ability. For all n records in the evaluation datasets, we simulate the test taker completing the adaptive version of the assessment using their recorded response to questions from the full assessment. The item selection optimization problem is solved at each iteration of the adaptive test, which we evaluate for all three solutions selected from the weight vectors.

The final results for each assessment dataset are presented in Table 3, which shows the average total number of questions (i.e., $\sum_{i=1}^{n} f_{1_i}(x)$ from every optimization problem solved at each CAT iteration), the average final SEM (i.e., the cut-off $f_2(x)$ of the final selected solution of the CAT iteration), and RMSE. Regardless of the weight vector used, the termination criteria, or SEM, controls the number of questions, as seen in the results where the total test length is reduced overall, but is similar across all W_i. However, this also indicates fewer CAT iterations are needed to reach a similar termination criteria. We find that the results for the dichotomous

Table 3: Final results for all datasets and weights.

Trait	W_i	Avg. Qs	Avg. SEM	RMSE
NPI	W_1	30.3548	0.2417	0.0888
	W_2	30.4558	0.2415	**0.0886**
	W_3	30.5651	0.2412	0.0925
MACH	W_1	2.6705	0.1509	1.1325
	W_2	2.6705	0.1509	**1.1319**
	W_3	2.6705	0.1509	1.1343
EQSQ-E	W_1	2.9204	0.1736	1.8286
	W_2	2.9204	0.1736	1.8291
	W_3	3.0000	0.1623	**1.7873**
EQSQ-S	W_1	3.1342	0.1674	1.7099
	W_2	3.1342	0.1674	1.7102
	W_3	3.0000	0.1646	**1.6904**
POL	W_1	29.1975	0.2115	0.2007
	W_2	29.2407	0.2120	**0.1873**
	W_3	30.2407	0.2091	0.1927

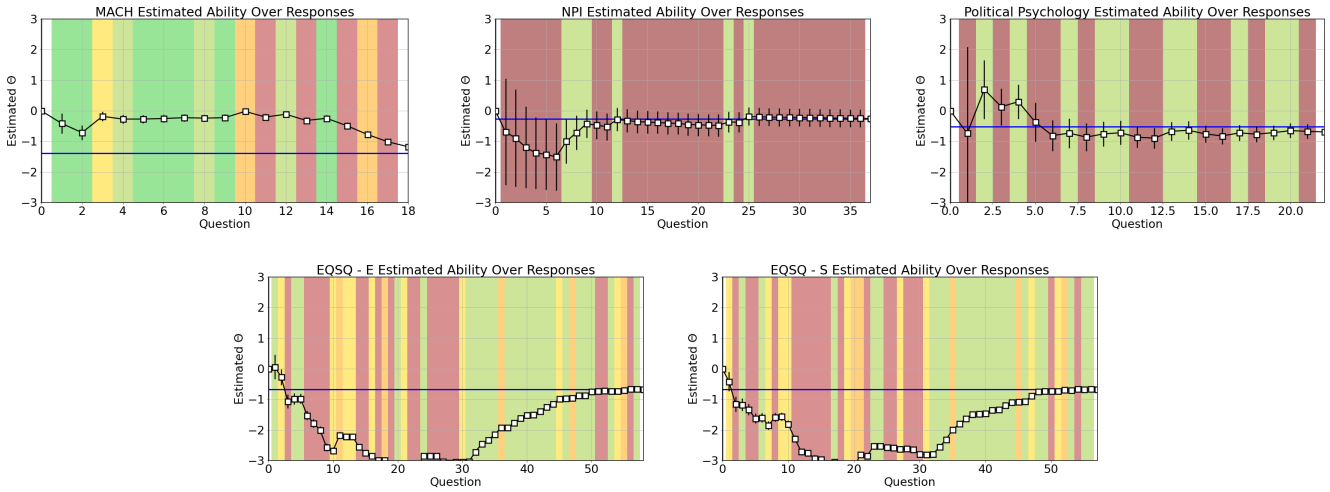

Figure 8: A sample of the new CAT using the MOEA approach. The black bar on each point indicates the error in ability estimation at each question. The blue bar in the middle of the plot indicates the actual ability from the full assessment.

assessments, NPI and POL, outperform polytomous assessments, MACH and EQSQ. This is due to the low SEM found in the polytomous Pareto-optimal solutions, resulting in a smaller adaptive assessment. This can be changed by updating the termination criteria of the CAT, or using a different weight vector, to increase assessment size. However, with the POL assessment, our results show the length of the test was reduced by nearly half the original size. Similarly, NPI results show a reduction by 25% of the full length assessment. Furthermore, the low RMSE values for both dichotomous assessments indicate highly accurate results, close to the original ability estimate by the full assessment.

Using W_2, we obtain the final sample results for the CAT, as shown in Figure 8, where each assessment is the reduced version of the original. The blue bar in the center of the plot indicates the actual estimated ability (as determined by the full assessment). We observe an estimated ability close to the actual value is achieved in fewer questions for NPI and POL, reducing the assessments by nearly 50%. MACH and EQSQ assessments used a lower CAT termination criteria while selecting the same solution from the Pareto-optimal set, and also had a slight reduction in test length. In addition, we observe that both converge to the correct assessment value, though only a slight reduction in test length for EQSQ and MACH. In comparison, the NPI and POL assessments were reduced by a greater portion of the original length.

6 CONCLUSION

We present a new multi-objective optimization method for item selection in computerized adaptive testing. The optimization problem minimizes test length, while also minimizing the standard error of measurement. An evolutionary multi-criterion optimization algorithm, NSGA-II, is used to solve the problem with a binary representation. Our approach is evaluated using four datasets, two with dichotomous items, and two with polytomous items. We use two different IRT models: the 2PL model for the dichotomous items datasets, and GRM for the polytomous items datasets. These models were calibrated using large datasets from previous studies, and evaluated using RMSE. The final Pareto-optimal solutions were validated by using single-objective optimization, and individual solutions were selected using pseudo weights. The new adaptive assessment was able to achieve 95% confidence in the estimated ability while reducing the size of each assessment. In particular, the NPI dataset test length was reduced by 25% of the original length, and the POL assessment results showed a 50% reduction. The results obtained indicate the MOEA approach is advantageous for item selection and should be further studied for CAT.

We propose three areas for future research. First, use of MOEAs for multidimensional CAT item selection, where multiple traits are measured simultaneously. This is challenging, since evaluation with large assessment datasets requires computational time and resources; our approach, focusing on unidimensional assessments, took approximately two days to run every simulated evaluation with all datasets. Second, investigating other CAT item selection methods (discussed in Section 2.1) as a substitute for the second objective in our MOEA approach may be beneficial to compare results in other studies. Last, it will be useful to assess our approach with real individuals. Though the datasets used contain responses from real individuals and are similarly valid and reliable to in-person studies [3], a study with the CAT in-person may provide new insight for survey researchers.

Our work presents the first MOEA for item selection in CAT. We find MOEAs can be beneficial in item selection with long assessments, capturing the trade-off between precision and test length. This will help survey researchers to efficiently design their studies with CAT, and improve assessments in many fields.

ACKNOWLEDGMENTS

This material is based upon work partly supported by the U.S. National Science Foundation under Grant No. 1936857.

REFERENCES

[1] [n.d.]. Open-Source Psychometrics Project. https://openpsychometrics.org/_rawdata/.

[2] Simon Baron-Cohen, Jennifer Richler, Dheraj Bisarya, Nhishanth Gurunathan, and Sally Wheelwright. 2003. The systemizing quotient: an investigation of adults with Asperger syndrome or high–functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358, 1430 (2003), 361–374.

[3] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. *PloS one* 10, 4 (2015), e0121595.

[4] Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. Introducing a framework to assess newly created questions with Natural Language Processing. *arXiv preprint arXiv:2004.13530* (2020).

[5] Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. R2DE: a NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 412–421.

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

[7] Denny Borsboom. 2006. The attack of the psychometricians. *Psychometrika* 71, 3 (2006), 425.

[8] Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. 2016. Item response theory. *Annual Review of Statistics and Its Application* 3 (2016), 297–321.

[9] Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yiying Chen, Haiping Ma, and Guoping Hu. 2019. DIRT: Deep Learning Enhanced Item Response Theory for Cognitive Diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2397–2400.

[10] Jaehwa Choi. 2020. AUTOMATIC ITEM GENERATION WITH MACHINE LEARNING TECHNIQUES. *Application of Artificial Intelligence to Assessment* (2020), 189.

[11] Richard Christie and Florence L Geis. 2013. *Studies in machiavellianism*. Academic Press.

[12] Kalyanmoy Deb. 2011. Multi-objective optimisation using evolutionary algorithms: an introduction. In *Multi-objective evolutionary optimisation for product design and manufacturing*. Springer, 3–34.

[13] 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.

[14] Kalyanmoy Deb, Karthik Sindhya, and Tatsuya Okabe. 2007. Self-adaptive simulated binary crossover for real-parameter optimization. In *Proceedings of the 9th annual conference on genetic and evolutionary computation*. 1187–1194.

[15] Mirta Galesic and Michael Bosnjak. 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly* 73, 2 (2009), 349–360.

[16] Kyung Chris Tyek Han. 2018. Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions* 15 (2018).

[17] Peter MC Harrison, Tom Collins, and Daniel Müllensiefen. 2017. Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports* 7, 1 (2017), 1–18.

[18] Mohammad Khajah, Robert V Lindsey, and Michael C Mozer. 2016. How deep is knowledge tracing? *arXiv preprint arXiv:1604.02416* (2016).

[19] Andrew S Lan and Richard G Baraniuk. 2016. A Contextual Bandits Framework for Personalized Learning Action Selection.. In *EDM*. 424–429.

[20] Xiao Li, Hanchen Xu, Jinming Zhang, and Hua-hua Chang. 2020. Deep Reinforcement Learning for Adaptive Learning Systems. *arXiv preprint arXiv:2004.08410* (2020).

[21] Rob R Meijer and Michael L Nering. 1999. Computerized adaptive testing: Overview and introduction.

[22] Jacob M Montgomery and Josh Cutler. 2013. Computerized adaptive testing for public opinion surveys. *Political Analysis* 21, 2 (2013), 172–192.

[23] Jacob M Montgomery and Erin L Rossiter. 2020. So many questions, so little time: Integrating adaptive inventories into public opinion research. *Journal of Survey Statistics and Methodology* 8, 4 (2020), 667–690.

[24] Remo Ostini and Michael L Nering. 2006. *Polytomous item response theory models*. Number 144. Sage.

[25] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949* (2019).

[26] Georg Rasch. 1960. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. (1960).

[27] Robert Raskin and Howard Terry. 1988. A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of personality and social psychology* 54, 5 (1988), 890.

[28] Dimitris Rizopoulos. 2006. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software* 17, 5 (2006), 1–25.

[29] Ryota Sekiya, Satoshi Oyama, and Masahito Kurihara. 2019. User-Adaptive Preparation of Mathematical Puzzles Using Item Response Theory and Deep Learning. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 530–537.

[30] Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics* 8 (2020), 247–263.

[31] Yanyan Sheng and Christopher K Wikle. 2007. Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement* 67, 6 (2007), 899–919.

[32] Baoxu Shi, Shan Li, Jaewon Yang, Mustafa Emre Kazdagli, and Qi He. 2020. Learning to Ask Screening Questions for Job Postings. *arXiv preprint arXiv:2004.14969* (2020).

[33] Masaki Uto and Yuto Uchida. 2020. Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory. In *International Conference on Artificial Intelligence in Education*. Springer, 334–339.

[34] Wim J Van der Linden. 2018. *Handbook of item response theory, three volume set*. CRC Press.

[35] Wim J van der Linden and Ronald K Hambleton. 2013. *Handbook of modern item response theory*. Springer Science & Business Media.

[36] Matthias von Davier. 2018. Automated item generation with recurrent neural networks. *psychometrika* 83, 4 (2018), 847–857.

[37] David J Weiss and G Gage Kingsbury. 1984. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement* 21, 4 (1984), 361–375.

[38] Mike Wu, Richard L Davis, Benjamin W Domingue, Chris Piech, and Noah Goodman. 2020. Variational Item Response Theory: Fast, Accurate, and Expressive. *arXiv preprint arXiv:2002.00276* (2020).

[39] Jiaying Xiao. 2019. Collaborative Filtering Item Selection Methods for On-the-Fly Assembled Multistage Adaptive Testing. (2019).

[40] Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 193–197.

[41] Chun-Kit Yeung. 2019. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738* (2019).

[42] Cristian Zanon, Claudio S Hutz, Hanwook Henry Yoo, and Ronald K Hambleton. 2016. An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica* 29, 1 (2016), 1–10.