Proceedings of Machine Learning Research 1:1–15, 2024 FM-EduAssess at NeurIPS 2024 Workshop

BanditCAT and AutoIRT: Machine Learning Approaches to Computerized Adaptive Testing and Item Calibration

James Sharpnack JAMES.SHARPNACK@DUOLINGO.COM
Kevin Hao KEVIN.HAO@DUOLINGO.COM
Phoebe Mulcaire PHOEBE.MULCAIRE@DUOLINGO.COM
Klinton Bicknell KLINTON.BICKNELL@DUOLINGO.COM
Geoff LaFlair GEOFF.LAFLAIR@DUOLINGO.COM
Kevin Yancey KEVIN.YANCEY@DUOLINGO.COM
Alina A. von Davier ALINA.VONDAVIER@DUOLINGO.COM
Duolingo, 5900 Penn Ave., Pittsburgh PA 15206

Abstract

In this paper, we present a complete framework for quickly calibrating and administering a robust large-scale computerized adaptive test (CAT) with a small number of responses. Calibration — learning item parameters in a test — is done using *AutoIRT*, a new method that uses automated machine learning (AutoML) in combination with item response theory (IRT), originally proposed in Sharpnack et al. (2024). AutoIRT trains a non-parametric AutoML grading model using item features, followed by an item-specific parametric model, which results in an explanatory IRT model. In our work, we use tabular AutoML tools (`AutoGluon.tabular`, Erickson et al. (2020)) along with BERT embeddings and linguistically motivated NLP features. In this framework, we use Bayesian updating to obtain test taker ability posterior distributions for administration and scoring.

For administration of our adaptive test, we propose the *BanditCAT* framework, a methodology motivated by casting the problem in the contextual bandit framework and utilizing item response theory (IRT). The key insight lies in defining the bandit reward as the Fisher information for the selected item, given the latent test taker ability (θ) from IRT assumptions. We use Thompson sampling to balance between exploring items with different psychometric characteristics and selecting highly discriminative items that give more precise information about θ. To control item exposure, we inject noise through an additional randomization step before computing the Fisher information. This framework was used to initially launch two new item types on the DET practice test using limited training data. We outline some reliability and exposure metrics for the 5 practice test experiments that utilized this framework.

1. Introduction

The goal of test assessment is to measure an abstract characteristic, like listening comprehension or speech production, and summarize the test taker's proficiency with a score. Computerized adaptive tests sequentially administer items (i.e., questions) according to an administration policy. Ideally, this policy maximizes the information gathered about the test taker's ability, e.g., by adjusting the difficulty of administered items based on previous test taker responses. Reported scores are designed to reflect the test taker's proficiency, which is achieved through proper calibration of the test items. In this paper, we introduce two methods for calibrating and deploying adaptive tests.

Calibration. Item response theory (IRT) models an item's psychometric characteristics, also known as item parameters, such as difficulty and discrimination, which are essential for adaptive testing and scoring. Item Response Theory (IRT) focuses on the measurement and evaluation of educational or psychological latent traits, that is, unobservable variables, such as ability, skill, or competence levels (Lord and Novick, 1950). However, traditional IRT calibration requires extensive data collection, typically gathered during an item piloting phase. There are three primary situations in which we might calibrate a proportion of an item bank, cold-start and jump-start piloting of new items, and warm-start recalibration of the entire item bank.

1. *Cold-start:* New pilot items are added without response data, and only item features are used to calibrate them, while re-calibrating the existing operational item bank.

2. *Jump-start:* A limited number of responses are collected for new pilot items in a piloting phase, while both pilot and operational items are calibrated.

3. *Warm-start:* The operational item bank is re-calibrated to reflect recent response data, often due to changes in the user interface, test taker population, or preparation materials.

When an entirely new item bank is to be launched, there is no existing data with which to calibrate the items, and typically we have a pre-launch phase where we gather responses with which to calibrate items. Explanatory IRT models refer to item response theory (IRT) models that use item features to predict item parameters, such as difficulty or discrimination, rather than treating each item independently. This enables calibrating tests in the cold and jump-start settings, by using item level features (such as NLP features) to predict the item parameters.

In this work, we outline AutoIRT, a new approach to calibrating item parameters using automated machine learning (AutoML). AutoML tools train machine learning models with automated tuning parameter selection, feature engineering, data processing, and task selection. AutoIRT leverages AutoML to train IRT models from test responses and item content, overcoming the need for manual tuning and supporting multi-modal input. While traditional AutoML models are not inherently interpretable, AutoIRT maintains compatibility with IRT models, the standard in psychometrics. AutoIRT was introduced in Sharpnack et al. (2024) and is the first use of AutoML to fit IRT models.

Adaptive Testing. A computerized adaptive test (CAT) is a psychological or educational test that is usually administered in a digital environment and it uses an iterative, algorithm-based approach to select and administer test items. In IRT-based CATs, items are chosen based on the TT's estimated ability level throughout the testing process, and the estimated ability is continuously updated after each item is responded to. Test developers aim to devise optimal sequences of items for each TT, so that the accuracy of the test is maximized, while keeping the test relatively short, the item exposure low, and the cheating rates by copying from other TT low. Hence, a CAT is a much more efficient test, being shorter while maintaining the accuracy of linear tests (Lord, 1980; Wainer, 2000). Although a CAT can be administered outside of an IRT framework, it is customary in psychometrics to use IRT for every part of the testing process: item bank calibration, item routing, and ability estimation.

In this work, we introduce a new approach to item selection in a CAT, called BanditCAT. This framework casts CAT in the contextual bandit framework where the rewards are the Fisher information of that item for the test taker's ability. The Fisher information is derived from the IRT item parameters that were estimated using AutoIRT. We use Thompson sampling along with an additional randomization step to control for exposure while still selecting informative items. We show the effectiveness of this system by outlining its use on the practice Duolingo English Test during the launch of two new item types, Y/N Vocabulary and Vocabulary-in-context. This preliminary report introduces the first method in the BanditCAT framework, but it does not exploit the full capability of this approach. A more complete follow-up work based on this framework is forthcoming.

1.1. Background on Computerized Adaptive Testing

Item Response Theory. IRT corresponds to a class of interpretable statistical models that is commonly used for the scoring and administration of psychological and educational assessments, including computerized adaptive tests (CATs) (Baker and Kim, 2004; Lord, 1980; Wainer, 2000). While there are multidimensional IRT models, in most operational applications the unidimensional IRT models are typically used. The unidimensional IRT framework makes the following assumptions: the test taker (TT) grade distribution depends on a single parameter θ (unidimensionality), the expected grade is increasing in ability (monotonicity), and grade measurements are independent given that the TT was administered the items (local independence) (Bock and Aitkin, 1981). IRT models the grades and ability θ for a given TT. Define G_i as the random grade if they were administered item i, then the item response function is the expectation $p_i(\theta) = \mathbb{E}[G_i|\theta]$ (by unidimensionality). A test session consists of a sequence of T administered items $\mathbf{I}_T = (I_1, I_2, \dots, I_T)$ and observed grades $\mathbf{G}_T = (G_1, \dots, G_T)$. The local independence assumption implies that these are both random with the following Markov structure: G_t is independent of $\mathbf{G}_{t-1}$ given I_t for $t = 1, \dots, T$ (Baker and Kim, 2004).

One important aspect of IRT models is that they are *interpretable*, which is achieved by specifying a model. The most commonly used IRT models are the parametric logistic models for binary grades (right or wrong response). In these models the grade probability, also known as the item response function (IRF), takes the form,

$$p(\theta; \phi_i) = \mathbb{P}\{G_t = 1|\theta, I_t = i\} = c_i + (1 - c_i) \cdot \sigma(a_i(\theta - d_i)), \tag{1}$$

where $\phi_i = (a_i, c_i, d_i)$ are the item parameters: slope, or discrimination, a_i, chance parameter, c_i, and difficulty, d_i (Lord, 1980; Fischer, 1973). The slope parameter is related to the efficiency of an item in terms of the information provided to distinguish among test takers. The chance parameter is part of the 3-parameter logistic IRT model, reflecting the probability of a correct response due to chance, and the difficulty parameter reflects the difficulty of the item for a particular TT population. There is a clear connection between the IRF and the item parameters. For example, at larger discrimination levels, the IRF is steeper, and a small increase in ability around that item's difficulty level will lead to a significant increase in the probability of a correct response (Baker and Kim, 2004; Fox, 2010).

Calibration with explanatory IRT Models. Without additional information about items, calibrating IRT models such as (1) requires hundreds of responses. In order to calibrate in the cold-start and jump-start settings, we need to use contextual information about the items, typically derived from the content of the item itself (e.g., lexical and semantic content). We assume throughout that each item has associated explanatory features, denoted as $\mathbf{x}_i \in \mathbb{R}^d$. These features may be derived from neural network embeddings such as BERT (Devlin et al., 2019) or CLIP (Radford et al., 2021), hand-crafted linguistic features, or other key characteristics of the item. The use of item features in explanatory IRT models has a long history. A prominent example is the Log-Linear Trait Model (LLTM) by Fischer (1973), which extends the Rasch model by using item features to predict item difficulty through a linear model (see also De Boeck and Wilson (2004)). More recent extensions leverage pretrained language models like BERT (Devlin et al., 2019) to calibrate items from limited response data. For instance, Benedetto et al. (2021) predicted item difficulty using a fine-tuned BERT model based on student responses to multiple-choice questions, while Byrd and Srivastava (2022) employed a similar approach to estimate both difficulty and discrimination parameters for 2PL IRT models from natural language questions. Their work incorporated linguistically motivated features such as semantic ambiguity, alongside contextual embeddings generated by BERT. Additionally, Reyes et al. (2023) used RNNs to predict item difficulty, which had been pre-fitted using the Rasch model. Most of these approaches focus on predicting item parameters that have been pre-estimated using a large item bank with substantial response data. Training such models requires an extensive item bank, with hundreds of responses per item to generate the necessary training data.

An alternative method is to train an explanatory IRT model by fitting the item parameters through a neural network (NNet), where the final layer represents the IRT model (1). This approach trains a NNet to predict scores from features and ability θ, while constraining the model to follow the IRT structure. BERT-LLTM was introduced in McCarthy et al. (2021) to train LLTM models using BERT embeddings alongside hand-crafted linguistic features. More recently, Yancey et al. (2024) proposed BERT-IRT, which fits a 2PL model by constraining the NNet to follow the form of (1) with $c_i = 0$ for all items i. In that work, they demonstrate that BERT-IRT can accurately train explanatory IRT models with only a few responses per item and apply it to English language proficiency assessments. They use a linear form for $\log(a_i)$ and d_i, which Sharpnack et al. (2024) enhanced by employing a fully non-parametric AutoML model. The model is trained using $\mathbf{x}_i$ and $\tilde{\theta}_s$ as inputs, with the predicted binary score as the output. In our experiments we used a use a proxy estimate for test taker ability, $\tilde{\theta}_s$, which is derived from scores on other item types (similar to the approach in Yancey et al. (2024)).

Computerized Adaptive Testing. There are several broad categories of CAT algorithms, such as maximum information selection (Birnbaum, 1968), maximum global information selection (Chang and Ying, 1996), fully Bayesian selection (van der Linden, 1998), and shadow testing (van der Linden, 2005). Most of these methods base their selection on the Fisher information of each item given our current estimate of θ for the model (1). For the 3PL model this is,

$$F_i(\theta) = \frac{(a_i \cdot (1 - c_i) \cdot p_2(\theta; a_i, d_i) \cdot (1 - p_2(\theta; a_i, d_i)))^2}{(c_i + (1 - c_i) \cdot p_2(\theta; a_i, d_i)) \cdot (1 - (c_i + (1 - c_i) \cdot p_2(\theta; a_i, d_i)))} \tag{2}$$

where

$$p_2(\theta; a_i, d_i) = \frac{1}{1 + \exp(-a_i(\theta - d_i))}, \tag{3}$$

is the 2PL IRF. The main exception to this are those that select based on the KL divergence (Chang and Ying, 1996). All of these methods require provisional estimates of θ, furthermore several require estimating the posterior distribution of θ given the responses observed thus far for the given test session. Luckily, because the IRT models are univariate, the posterior can be estimated without trouble by approximating distributions as point masses on a discrete grid of θ values. In multi-dimensional models, this approach breaks down and more sophisticated Bayesian updating methods are needed (Mulder and van der Linden, 2010).

Directly selecting items that maximize this information measure can result in poor exposure rates, namely, a handful of items can be selected far too frequently. Several methods have been proposed for exposure control including item cloning (van der Linden and Veldkamp, 2000) and the Sympson-Hetter method (Sympson and Hetter, 1985). While the latter tracks the current exposure rates and guides them toward a target, the former uses randomization to balance exposure by randomly selecting items from the top k most informative. The randomization approach has a practical advantage in production systems because they are not dependent on any global statistics that need to be updated, and so can be deployed to servers independently who do not need a common cache.

Thompson Sampling and Contextual Bandits. A stochastic contextual bandit problem models a repeated interaction between a player and an environment. In each round, the player is presented with information about K actions (called arms), typically represented by d-dimensional feature vectors. The player must select an arm based on past observations. Only the reward from the selected arm is revealed, and the relationship between the rewards and features is governed by a linear model. The objective of the player is to maximize cumulative reward over T rounds. Over the past few decades, bandit algorithms have gained widespread use in various real-world applications, such as recommender systems (Li et al., 2010), online advertising (Schwartz et al., 2017), and clinical trials (Woodroofe, 1979).

Classic stochastic linear contextual bandit algorithms, such as Linear Upper Confidence Bound (LinUCB) (Li et al., 2010) and Linear Thompson Sampling (LinTS) (Agrawal and Goyal, 2013), have been shown to achieve nearly optimal total reward, given some modeling assumptions. Like most bandit algorithms, the objective is to begin by exploring the space of actions to get a sense of which are profitable, and then exploiting the actions that will give the highest reward. Thompson sampling accomplishes this by modeling reward as a function of the context vectors and latent parameters. By tracking the posterior distribution of the latent parameters, it draws realizations of these from the posterior and then selecting actions that maximize expected reward evaluated at that draw. Translating this to the CAT setting, the 'player' is the CAT, 'environment' is the test taker, 'actions' are the items administered, 'context' is based on the previous responses of the test taker and information about the items themselves, 'reward' is the information gained about test taker ability.

2. Item Calibration with AutoIRT

To harness the flexibility and performance of AutoML, we begin by fitting a grade classifier using the ability parameter θ and d-dimensional item features $\mathbf{x}_i \in \mathbb{R}^d$ as inputs. In our experiments, we employ a stacked ensemble of models—random forests, LightGBM, XGBoost, and CATBoost—implemented through the AutoGluon-tabular Python package Erickson et al. (2020). This package was chosen for its strong performance in tabular data benchmarks Gijsbers et al. (2024). Although we focus on tabular models since we have already engineered features like BERT embeddings, AutoML supports multimodal input. We hope to eventually apply AutoIRT directly to raw item content. The item ID is passed to the AutoML predictor as a feature, similar to the use of random effects in traditional models.

Let $\hat{p}(\theta; \mathbf{x}_i)$ represent the AutoML-predicted probability of a correct response for a test taker with ability parameter θ and an item with features $\mathbf{x}_i$. Next, we extract a more interpretable IRT model by projecting the AutoML model onto the closest IRT model using a least squares approach. Specifically, we minimize the following loss function to estimate item parameters:

$$L(\phi) = \sum_{i \in \mathcal{R}} \sum_{\theta \in \Theta} (p(\theta; \phi_i) - \hat{p}(\theta; \mathbf{x}_i))^2,$$

where Θ is a regular grid of θ values. This ensures that Θ covers most of the probability mass of the distribution of θs. In the case of the 3PL model (1), we either will allow the chance parameter c_i to vary, or be fixed (typically at $c_i = 0$, the 2PL model). In Sharpnack et al. (2024), an extension of this model was proposed which uses the above method as the M-step in a Monte Carlo EM algorithm. This allowed them to jointly learn θ and the item parameters ϕ_i. One restriction is that the chance parameters and θ cannot be jointly learned due to a lack of identifiability (Baker and Kim, 2004), so that approach required fixing chance c_i. In our setting, we initialize θ to be a weighted combination of other item type scores, resulting in a proxy ability. This approach is similar to what was taken in Yancey et al. (2024).

3. BanditCAT Framework

3.1. Contextual Bandit Interpretation of CATs

In this section, we will cast CAT administration as a contextual bandit problem. Each arm in the bandit setting corresponds to an item I_t and the action is the administration of the item at round t of T rounds. Each item has context, which are the item features, $x_i \in \mathbb{R}^d$. The reward is the Fisher information for that item $F_{I_t}(\theta)$ for that test taker's true θ. This reward is unobserved because we do not have access to the true θ, but we are able to model the reward via our provisional estimate of θ (or the current posterior) and our calibrated item parameters. The grades enter into the picture via our provisional estimate of θ, and our understanding of how informative an item is for a test taker's ability is a direct product of the observed grades, G_{I_t}.

One of the advantages of using the Fisher information is that it directly measures the information content of a response to that item. At the end of a test session we can evaluate how effective our item selections were by measuring the information content of the data that we gathered about the test taker. Specifically, the total reward for that session is the Fisher information of all data gathered thus far,

$$\bar{F}_T(\theta) := \mathbb{E}\left[\left(\sum_{t=1}^{T} \frac{\partial}{\partial \theta} \log f(G_{I_t}|\theta)\right)^2\right] = \sum_{t=1}^{T} F_{I_t}(\theta). \quad (4)$$

The above follows from the fact that the score (gradient of log likelihood) at the true θ has mean 0 and the local independence assumption. It should be noted that we are conditioning on the items that were administered, $\mathbf{I}_n$, in that session. We know by the Cramer-Rao lower bound that the variance of any unbiased estimator for θ is at least $1/F_T(\theta)$, under regularity conditions, that are satisfied in this case.

For the 2PL model, there is a clear relationship between the item parameters and the Fisher information, namely the height is $a_i^2/4$, regardless of d_i. This is not true of the 3PL model, since the chance parameter makes the maximum information point deviate from d_i. The height is then a complex function of the item parameters, which will not be amenable to our exposure control randomization method. To this end, we approximate the 3PL Fisher information with a Gaussian kernel,

$$F_i(\theta; \mu_i, \nu_i, h_i) = h_i \cdot \exp\left(-\frac{(\theta - \mu_i)^2}{2\nu_i^2}\right), \tag{5}$$

where μ_i is the center parameter, h_i is the height, and ν_i is the bandwidth parameter. For a given set of item parameters a_i, c_i, d_i we obtain μ_i, h_i, ν_i by moment matching. For the 2PL model, we use (2) directly, which simplifies greatly.

3.2. Thompson Sampling with Exposure Control

In this section, we introduce an initial algorithm for item selection under the BanditCAT framework, which we call BanditCAT V1 to indicate that it is the first incarnation of this approach. In our setting the reward $F_{I_t}(\theta)$ is unobserved, however we can build a model for it via the posterior $\pi(\theta|G_{\mathbf{I_t}})$. To utilize the bandit framework, we let the reward model be the following,

$$\hat{r}_i(\theta) = F_i(\theta; \hat{\phi}_i), \tag{6}$$

where $\hat{\phi}_i$ is the predicted item parameters $(\hat{a}_i, \hat{c}_i, \hat{d}_i)$ from AutoIRT. In the case of the Gaussian kernel Fisher information (5) we first convert the item parameters $\hat{\phi}_i$ into kernel parameters $(\hat{h}_i, \hat{\mu}_i, \hat{\nu}_i)$. In order to utilize the full contextual bandit framework, the item parameters $\hat{\phi}_i$ should be drawn from a distribution, which we reserve for a future work.

We want to control the frequency of arm selection (i.e., exposure) for Thompson sampling in bandits, and we accomplish this using randomization in the item parameters when calculating $\hat{r}_i(\theta)$. Specifically, for the 2PL model, we sample the discrimination parameter,

$$\tilde{a}_i \sim \text{Gamma}(\hat{a}_i/\gamma, \gamma), \tag{7}$$

where $\gamma > 0$ is a global scale parameter (the mean for the Gamma is $\hat{a}_i$). For the 3PL model we use the Fisher information approximation, (5), and instead add randomness to the height parameter,

$$\tilde{h}_i \sim \text{Gamma}(\hat{h}_i/\gamma, \gamma). \tag{8}$$

These randomization approaches were developed through additional experimentation and simulation. In our method, we draw θ from the posterior distribution given the observed grades. However, we additionally allow for multiple θ_k's to be drawn and then we average the estimated rewards for each of these. The purpose of this is to interpolate between traditional Thompson sampling and fully Bayesian item selection (van der Linden, 1998). Specifically, the expectation of $F_i(\theta)$ over the posterior is the Bayesian information criteria, so this is approximated by the average with enough Monte Carlo samples of θ.

3.3. BanditCAT V1

Input: eligible item bank $\mathcal{I}$, fitted item parameters $\hat{\phi}_i$ (from AutoIRT), prior distribution over θ, π_1, and exposure control parameter $\gamma > 0$.

At each round, $t = 1, \ldots, T$,

1. For $k = 1, \ldots, K$, independently draw $\theta_k \in \Theta$ from $\pi_t(.)$.
2. For each test item, $i \in \mathcal{I}$, draw information parameters,
   (a) For the 3PL model, draw $(\tilde{h}_i, \hat{\mu}_i, \hat{\nu}_i)$ from (8) to obtain $\tilde{r}_i(\theta_k), k = 1, \ldots, K$, according to (5).
   (b) For the 2PL model, draw $(\tilde{a}_i, \hat{d}_i)$ from (7) to obtain $\tilde{r}_i(\theta_k), k = 1, \ldots, K$, according to (2).
3. Select item $I_t \in \mathcal{I}$ that maximizes $\frac{1}{K} \sum_{k=1}^{K} \tilde{r}_i(\theta_k)$.
4. Observe the grade G_{I_t} for that test and update posterior $\pi_t(.) = \pi(.|G_{I_1}, \ldots, G_{I_t})$.

Finally, we return the posterior mean $\mathbb{E}_{\pi_T}[\theta]$ as the score for that item type.

Aside from its statistical soundness there are some advantages to this method. Particularly, it can be deployed independently to any server without any shared cache or memory, making server autoscaling quite simple. Computationally the method is quite simple, with the most expensive operation being the very fast posterior calculation. In the instance that there are additional eligibility criteria, then BanditCAT V1 will reduce the eligible item bank prior to step (2). For multiple item types, this method is applied in sequence for each item type. Multidimensional variants of BanditCAT are in development as well as a complete Thompson sampling approach that samples the item parameters as well.

3.4. CAT Simulation

For model selection and tuning config parameters (most notably the exposure control parameter γ), we need to be able to simulate what will happen when we apply our administration algorithm to new test takers. We use the following **nearest neighbor matching algorithm** to simulate a new session for single text vocab.

Input: Test sessions with proxy θ's derived as a weighted average of scores from item types other than those being modified and responses for the item types in question.

For a sample of real tests, calculate their proxy θs (call this S_0) and for each test do the following:

1. Administer the first item, I_1, for a session.
2. For each round $t = 1, \ldots, T$,
   (a) Find the historical session that was administered item I_t that has the closest proxy θ to the target S_0 (nearest neighbor matching).
   (b) Let the grade G_t for the simulated session be the grade for the matched historical session on item I_t.
   (c) Update posterior and select next item I_{t+1}.

3. Score this simulated session as if it were real

We can use these simulated sessions to evaluate our desired metrics, which is typically the exposure rates so that we can hit our target.

3.5. Assessing Vocabulary in the Duolingo English Test

In this work, we study the calibration and administration of vocabulary item types in the Duolingo English Test (DET). We focus on two item types that are at the start of the DET which assess vocabulary knowledge. These items in combination assess the form, meaning, and use of English words.

Y/N Vocabulary and Vocab-in-Context item types. We focus on the first two item types in the DET (V8): yes/no vocabulary (Y/N Vocab) and vocabulary in context (ViC). The practice test includes 3290 Y/N Vocab items and 585 ViC items, with 18 Y/N Vocab and 9 ViC items administered per test session in succession. Test takers have 5 seconds to respond to Y/N Vocab items and 20 seconds for ViC items. Y/N Vocab requires the test taker to identify whether a word is real or fake (with fake words generated by an RNN). ViC is a fill-in-the-blank task where the test taker completes a word within a sentence. For example, a Y/N Vocab item might ask if "newbacal" is a real word, while a ViC item might ask for the missing word in "I'm sorry for the inter_ _ _ _ _ _ _ _, but could you explain that last part again?" Additional details are provided in Sharpnack et al. (2024), and practice tests can be taken for free at `http://englishtest.duolingo.com`.

Item Features. Throughout we use a combination of linguistic and features from a foundation model (RoBERTa). For each item, we calculate a set of features based on its content, using the Corpus of Contemporary American English (COCA), Davies (2008), for corpus-based features. In Y/N Vocab items, which consist of a single word, features include a binary indicator for whether the word is real, its length, and binary indicators for its presence in CEFR-specific wordlists. We also consider log frequency, log frequency-rank, capitalization frequency, and n-gram features from COCA. For ViC items, we use surface features such as the number of missing characters and vowel proportion in the missing part, along with log frequencies from COCA and its sub-corpora. We include RoBERTa embeddings of the masked passage (reduced to 10 dimensions using PCA), sentence-level log frequencies, the normalized position of the damaged word, and the conditional probability of the correct word based on visible letters and word length. The use of foundation model features is indispensable to our calibration performance.

Administration of Vocabulary Items. The control condition in our experiments is the previous version of the DET practice test, which includes Y/N Vocab items but no ViC items. The control Y/N vocab items are arranged in a card format consisting of 18 words per card (roughly half real and half fake). The test taker is instructed to identify the real words within the card, and thoughout the test there are 4 such cards, for a total of 72 Y/N Vocab items. The card format has the disadvantage that it doesn't enable strong adaptivity, since the words displayed on the card are predetermined.

The administration algorithm for treatment conditions is the following. 18 Y/N Vocab items and 9 ViC items are administered in sequence at the beginning of DET test sessions. For each Y/N Vocab administration event, we draw a Bernoulli(1/2) to determine if the word displayed is real or fake. Then the real or fake item is selected using BanditCAT V1 from among the eligible items. Only items that have not yet been administered are eligible. The session continues with 9 ViC items with 20 second time limits after an instructional screen.

4. Experimental Results

4.1. DET Experiments

Table 1 summarizes a series of experiments conducted for the DET practice tests conducted between 2023-01-23 and 2023-02-19. The primary purpose of these experiments are to converge on a final candidate for DET V8. While these experiments were important steps along the way to obtaining this launch candidate, none of the treatment conditions listed here are operational on the DET. None of these experiments or their metrics are reflective of the certified DET. All treatment conditions listed here use AutoIRT + BanditCAT V1 as explained above. The experiments are grouped by experimental blocks (E1 to E5) and contain control and treatment conditions (e.g., T1, T2, T3, etc.). The control conditions (labeled as "C") across different experimental blocks use 4x Y/N (yes/no) vocab items, while the treatment conditions only vary in how they administer other aspects of the test such as other item types Some experiments, such as those marked with "*", denote adjusted calibrations of the AutoIRT method, with changes to the model or the downstream administration from a prior condition. Experiment E5 includes ViC items in addition to the Y/N Vocab items because this is the first experiment that used AutoIRT + BanditCAT V1 for ViC items. In these treatment conditions, the ViC administration was not dependent on the responses to the Y/N Vocab items. Each condition has a different number of participants (N) ranging from roughly 4,000 to over 14,000, indicating robust data collection efforts across different time spans in early 2023. In all of these cases, the items were calibrated in the jump-start setting, with a limited amount of response data for both of these item types. This preliminary response data was gathered from practice test sessions that were administered a treatment condition in a prior experiments (that we will not be discussing here). In all experiments, we set the exposure control parameter using the aforementioned nearest neighbor matching to hit a reasonable target.

4.2. Evaluation Metrics

Retest reliability (RR) is a common performance measure, defined as the Pearson correlation between scores when the same user takes the test twice. In classical test theory, the observed score is seen as the true score plus noise. The standard error of measurement (SEM), which estimates how much a score deviates from the true score, is related to RR by the formula:

$$S_E = S_X\sqrt{1 - RR},$$

where S_X is the population score standard deviation, and S_E is the error standard deviation. For example, increasing RR from 0.5 to 0.6 reduces SEM by 11.8%. In practice tests, the reliability is typically smaller than in certified tests where the test taker's motivation is higher. Another common metric for a item type score (such as the Y/N Vocab score) is the correlation between that item types score and the overall score for the test. The DET consists of 14 task types ranging from interactive listening and reading to dictation (Cardwell et al., 2022). We also consider the maximum exposure rates, that is over all of the test

Exper.	**Cond.**	**Min Date**	**Max Date**	**N**	**Description**
E1	C	01-23	01-26	6304	Control, 4x Y/N Vocab Items
E1	T1	01-23	01-26	6535	Y/N Vocab uses AutoIRT and BanditCAT V1
E1	T2	01-23	01-26	6050	Y/N Vocab uses AutoIRT and BanditCAT V1, different downstream admin. from T1
E2	C	01-26	01-31	10560	Control, 4x Y/N Vocab Items, no ViC
E2	T1*	01-26	01-31	10311	Adjusted calibration of T1
E2	T2*	01-26	01-31	10292	Adjusted calibration of T2
E3	C	01-31	02-05	14086	Control, 4x Y/N Vocab Items, no ViC
E3	T3	01-31	02-05	13207	Same as T2* above with different downstream admin.
E3	T2*	01-31	02-05	13855	Same as T2* above
E4	C	02-12	02-16	7498	Control, 4x Y/N Vocab Items, no ViC
E4	T4	02-12	02-16	7021	Same as T2* above with different downstream admin.
E4	T2*	02-12	02-16	14118	Same as T2* above
E5	C	02-16	02-19	4834	Control, 4x Y/N Vocab Items, no ViC
E5	T5	02-16	02-19	4301	Y/N Vocab and ViC uses AutoIRT and BanditCAT V1
E5	T6	02-16	02-19	8638	Y/N Vocab and ViC uses AutoIRT and BanditCAT V1, different downstream admin. from T5

Table 1: A description of the DET practice test experiments and their conditions.

sessions, what is the maximum frequency of an item being selected (f_i) as a ratio of the total number of administration events for items of that type.

4.3. Results Discussion

Table 2 gives the results of our experiments in the DET practice test. When it comes to the Y/N Vocab and ViC sections, the treatment conditions do not differ substantially with the exception of going from $T1, T2$ to $T1^*, T2^*$. In that case, the recalibration improved scoring and administration thus increasing the reliability metrics. In general, the control condition has significantly higher reliabilities and score correlations, but this is due to the fact that it has 4x the number of items as the treatment. We see that the loss in reliability for Y/N vocab when we reduce the number of items from 72 to 18 does not result in a similar loss of reliability (which we would expect to roughly half). Similarly, the drop in score correlations is not as significant as we might expect given the dramatic reduction in the number of items. We see that we were very effectively able to control exposure, since we did not see a significant change from the control condition.

While there is no control comparison for the ViC item type, we see that with 9 items we get a reliability of 0.66, which exceeds that for the 18 Y/N vocab items. It should be noted

Experiment	Cond.	Task	RR	Score Corr.	Max exposure
E1	C	Y/N Vocab	0.689	0.813	1.00%
E1	T1	Y/N Vocab	0.535	0.723	0.56%
E1	T2	Y/N Vocab	0.550	0.733	0.57%
E2	C	Y/N Vocab	**0.724**	0.824	0.98%
E2	T1*	Y/N Vocab	0.579	0.752	0.65%
E2	T2*	Y/N Vocab	**0.587**	0.753	0.67%
E3	C	Y/N Vocab	0.722	**0.825**	0.95%
E3	T3	Y/N Vocab	0.571	0.743	0.64%
E3	T2*	Y/N Vocab	0.567	0.758	0.61%
E4	C	Y/N Vocab	0.713	0.821	0.91%
E4	T4	Y/N Vocab	0.538	0.748	0.74%
E4	T2*	Y/N Vocab	0.543	0.741	0.74%
E5	C	Y/N Vocab	0.713	0.802	0.97%
E5	T5	Y/N Vocab	0.535	0.738	0.66%
E5	T6	Y/N Vocab	0.552	**0.759**	0.64%
E5	T5	ViC	0.659	0.737	1.10%
E5	T6	ViC	**0.661**	**0.742**	1.01%

Table 2: Reliability, validity, and exposure results of DET practice test experiments. We have bolded the maximum metrics for Y/N Vocab, ViC tasks, and control (for Y/N Vocab). All metrics listed are not reflective of the DET certified test or the current version of the practice test.

that the combination of these two item type scores will get significantly higher reliability (0.8 for E5-T6 for the average scores). The original time allotted to vocabulary items in the control condition is 4 minutes. The time allotted for Y/N Vocab and ViC combined is 4:30, yet there is a very substantial total reliability increase.

5. Conclusions

This work demonstrates a general approach to calibrating and administering an adaptive test with minimal responses and limited modeling effort. It outlines the AutoIRT method for obtaining item parameter estimates from complex item features, as well as foundation model embeddings. Specifically, we use the RoBERTa embeddings for items that contain passages (vocabulary-in-context). We also introduce the BanditCAT framework, which casts computerized adaptive tests in the contextual bandit setting. We provide an algorithm that is a first attempt at using this framework, BanditCAT V1 — a Thompson sampling approach to test administration.

While this paper does introduce the BanditCAT framework, BanditCAT V1 is just the first algorithm that utilizes this approach. However, this algorithm does not address a few major issues that plague adaptive testing. First, it is not a complete Thompson sampling algorithm because it does not also sample the item parameters from a posterior distribution. Because this is another critical source of uncertainty in the reward model (Fisher information), this would improve the balance between exploration and exploitation. Second, the method for exposure control is specific to the 2PL and 3PL IRT models, a general purpose exposure control would extend this approach to other IRT models. Third, this model only supports unidimensional θ, and to extend this problem to multidimensional θ would require additional modifications. This work is a precursor to a more complete treatment that addresses these issues.

References

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Frank B. Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC Press, Boca Raton, FL, 2nd edition, 2004. ISBN 978-0824758257.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–157, Online, 2021. Association for Computational Linguistics.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical theories of mental test scores*, pages 395–479. Addison-Wesley, Reading, MA, 1968.

R. Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

Matthew Byrd and Shashank Srivastava. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130, Dublin, Ireland, 2022. Association for Computational Linguistics.

Ramsey Cardwell, Geoffrey T LaFlair, and Burr Settles. Duolingo english test: technical manual. *Duolingo Research Report*, 2022.

Hua-Hua Chang and Zhiliang Ying. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229, 1996.

Mark Davies. Word frequency data from The Corpus of Contemporary American English (COCA). https://www.wordfrequency.info, 2008.

Paul De Boeck and Mark Wilson. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, NY, 2004.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Gerhard H. Fischer. *Log Linear Trait Models: An Approach to Item Analysis and Test Construction.* Psychometric Society, Chicago, IL, 1973.

Jean-Paul Fox. *Bayesian item response modeling: Theory and applications.* Springer, 2010.

Pieter Gijsbers, Marcos LP Bueno, Stefan Coors, Erin LeDell, Sébastien Poirier, Janek Thomas, Bernd Bischl, and Joaquin Vanschoren. Amlb: an automl benchmark. *Journal of Machine Learning Research*, 25(101):1–65, 2024.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems.* Routledge, Hillsdale, NJ, 1980.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA, 1950.

Arya D McCarthy, Kevin P Yancey, Geoffrey T LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 883–899, 2021.

Joris Mulder and Wim J. van der Linden. Multidimensional adaptive testing with kullback-leibler information item selection. In Wim J. van der Linden and Cees A.W. Glas, editors, *Elements of Adaptive Testing*, pages 169–192. Springer, New York, NY, 2010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Diego Reyes, Abelino Jimenez, Pablo Dartnell, Séverin Lions, and Sebastián Ríos. Multiple-choice questions difficulty prediction with neural networks. In *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*, pages 11–22. Springer, 2023.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

James Sharpnack, Phoebe Mulcaire, Klinton Bicknell, Geoff LaFlair, and Kevin Yancey. Autoirt: Calibrating item response theory models with automated machine learning. *arXiv preprint arXiv:2409.08823*, 2024.

James B. Sympson and Robert D. Hetter. Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th Annual Meeting of the Military Testing Association*, pages 973–977, 1985.

Wim J. van der Linden. Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2):201–216, 1998.

Wim J. van der Linden. A hybrid method for constrained adaptive testing. In Wim J. van der Linden and Cees A.W. Glas, editors, *Elements of Adaptive Testing*, pages 269–288. Springer, New York, NY, 2005.

Wim J. van der Linden and Bernard P. Veldkamp. Adaptive testing with item cloning. *Applied Psychological Measurement*, 24(2):129–150, 2000.

Howard Wainer. *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2000.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Kevin P Yancey, Andrew Runge, Geoffrey Laflair, and Phoebe Mulcaire. Bert-irt: Accelerating item piloting with bert embeddings and explainable irt models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438, 2024.