Title

Field-Testing Multiple-Choice Questions With AI Examinees: English Grammar Items.

Permalink
https://escholarship.org/uc/item/6418q27g

Author
Maeda, Hotaka

Publication Date
2024-10-03

DOI
10.1177/00131644241281053

Peer reviewed

Educational and Psychological
Measurement
1–24

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644241281053
journals.sagepub.com/home/epm
Sage

Field-Testing Multiple-Choice Questions With AI Examinees: English Grammar Items

Hotaka Maeda[1,2]

Abstract

Field-testing is an essential yet often resource-intensive step in the development of high-quality educational assessments. I introduce an innovative method for field-testing newly written exam items by substituting human examinees with artificially intelligent (AI) examinees. The proposed approach is demonstrated using 466 four-option multiple-choice English grammar questions. Pre-trained transformer language models are fine-tuned based on the 2-parameter logistic (2PL) item response model to respond like human test-takers. Each AI examinee is associated with a latent ability θ, and the item text is used to predict response selection probabilities for each of the four response options. For the best modeling approach identified, the overall correlation between the true and predicted 2PL correct response probabilities was .82 (bias = 0.00, root mean squared error = 0.18). The study results were promising, showing that item response data generated from AI can be used to calculate item proportion correct, item discrimination, conduct item calibration with anchors, distractor analysis, dimensionality analysis, and latent trait scoring. However, the proposed approach did not achieve the level of accuracy obtainable with human examinee response data. If further refined, potential resource savings in transitioning from human to AI field-testing could be enormous. AI could shorten the field-testing timeline, prevent examinees from seeing low-quality field-test items in real exams, shorten test lengths, eliminate test security, item exposure, and sample size concerns, reduce overall cost, and help expand the item bank. Example Python code from this study is available on Github: https://github.com/hotakamaeda/ai_field_testing1

[1]Smarter Balanced, Santa Cruz, CA, USA
[2]University of California–Santa Cruz, USA

Corresponding Author:
Hotaka Maeda, Smarter Balanced, 1156 High St, Santa Cruz, CA 95064, USA.
Email: hotaka.maeda@gmail.com

Keywords

natural language processing, pretesting, artificial intelligence, BERT, question difficulty prediction

Introduction

Assessments are integral to education, providing valuable feedback on student learning progress and guiding enhancements in teaching and curriculum design. To develop and maintain high-quality educational assessments, field-testing, or pretesting, is necessary to evaluate the quality of test items before they are used for scoring. In many large-scale assessments, field-testing involves asking a sample of examinees from the target population to complete unscored field-test items, often embedded among operational scored items. The quality of the field-test items is evaluated using a variety of statistical techniques, and is calibrated for future scoring. However, traditional field-testing methods are often resource-intensive, time-consuming, and can raise item exposure concerns (AlKhuzaey et al., 2023; Hsu et al., 2018; Jiao & Lissitz, 2020). In many cases, thousands of examinee test data are needed to conduct sufficiently accurate analyses (Morizot et al., 2007).

The literature provides successful cases of using collateral information to improve item parameter calibration efficiency (Mislevy, 1988; Wang & Jiao, 2011). But some researchers took this idea further by attempting to bypass the field-testing process altogether. These studies present models that predict item difficulty from various item text features, such as semantic and syntactic complexity, word and sentence lengths and counts, word embeddings, and readability indices (see AlKhuzaey et al., 2023; Benedetto et al., 2023). While some methods relied on expert judgment (Beinborn et al., 2014; Choi & Moon, 2020; Loukina et al., 2016; Settles et al., 2020), these subjective approaches often suffered from poor inter-rater reliability (i.e., consistency between multiple judges) and limited reproducibility (AlKhuzaey et al., 2023; Conejo et al., 2014). Other approaches relied on machine-driven natural language processing (NLP) techniques to predict item difficulty and/or discrimination (Benedetto et al., 2020a, 2020b, 2021; Yaneva et al., 2019; Zhou & Tao, 2020). However, the accuracy of these prediction methods remains limited, and merely estimating item difficulty and discrimination fails to capture the full scope of traditional field-testing. In addition, these models often do not replicate the complete human test-taking experience, such as by excluding the distractor options in their predictions (e.g., Benedetto, 2023; Benedetto et al., 2021).

An alternative approach to estimating item difficulty has been developed by Lalor et al. (2019) using artificial intelligence (AI). The researchers corrupted (i.e., altered or damaged) a random percentage of sentiment analysis data to intentionally create 1,000 sets of training data with varying levels of quality. Training deep learning models on these data resulted in 1,000 models with varied levels of accuracy in completing sentiment analysis tasks. They then used these models to generate response data and fitted item response theory (IRT) models to estimate the difficulty of completing each NLP task (also see Rodriguez et al., 2021). Others have used this approach to estimate item discrimination parameters as well (Byrd & Srivastava, 2022).

Although originally developed in the area of computer science, Lalor et al.'s (2019) approach could be adapted and applied to field-testing educational assessment items. This adaptation could lead to the development of *AI field-testing* that replaces human examinees with AI examinees, while preserving the natural flow and structure of traditional human field-testing. Compared to past item difficulty and discrimination prediction methods (e.g., AlKhuzaey et al., 2023), there is one key major advantage. AI examinees have the potential to simulate the test-taking process and generate item response data that closely mimic patterns observed in human responses. If successful, these AI-generated data could be analyzed in the same ways as traditional field-test data, offering applications far beyond merely predicting item difficulty. Therefore, I propose a method of training transformer large-language models (LLMs) by integrating them with IRT (Lord, 1980) to generate human-like item responses, which could then be used for field-test analyses. I will briefly introduce some backgrounds on LLMs and IRT in the upcoming sections. Then, I will evaluate the methods using data from a large-scale English language assessment program. This article builds on the preliminary work by Maeda (2023).

Background on Transformer Language Models

Introduced by Vaswani et al. (2017), the transformer neural network revolutionized NLP. Unlike recurrent neural networks, transformers utilize attention mechanisms, allowing them to consider all positions in an input sequence simultaneously. This parallelization significantly accelerated training and made transformers highly scalable.

Transformer language models like Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018) are pre-trained on large amounts of corpora. Since they are available open-access, users can download pre-trained models and customize them on specific downstream tasks. Input text is first tokenized and converted into word embeddings, which results in a numeric vector for each word or part of a word. These numerical representations of text are passed through the encoder layers of the transformers neural network, where the final layer is often fine-tuned to complete a specific NLP task. These include language translation, text summarization, question answering, sentiment analysis, or regression.

Transformers have become the backbone of contemporary NLP models, facilitating nuanced understanding of context and semantic relationships within language data, and achieving state-of-the-art performance on benchmarks. They excel in capturing the contextual meaning of long texts and can differentiate between various definitions of the same word (see Devlin et al., 2018).

In this article, the pre-trained DeBERTa-v3-large transformer language model (He et al., 2021) was used. DeBERTa (He et al., 2020) is an enhanced variant of the BERT (Devlin et al., 2018) and RoBERTa models (Liu et al., 2019). Each word in DeBERTa is associated with two vectors that independently represent its content and positioning. The attention weights between words are also calculated separately for their content and positioning. This allows, for example, for DeBERTa to learn that word ‘‘artificial’’ is highly associated with ‘‘intelligence’’ when they occur next to each other, unlike when there are other words between them. More recently, the training efficiency of DeBERTa was enhanced by utilizing replaced token detection techniques rather than masked language modeling. The resulting model was named DeBERTa-v3, which has 304 million parameters (He et al., 2021).

Background on Psychometric Theories

IRT is widely used for measuring latent abilities in educational assessment (Lord, 1980). Let $X_{ij} = 1$ and $X_{ij} = 0$ denote a correct and an incorrect response on a dichotomous item i for examinee j, respectively. A commonly used unidimensional IRT model is the 2-parameter logistic (2PL) model (Birnbaum, 1968), which expresses the probability of correct response as

$$P_{ij} = P(X_{ij} = 1|\theta_j) = \frac{\exp[1.7a_i(\theta_j - b_i)]}{1 + \exp[1.7a_i(\theta_j - b_i)]} \tag{1}$$

where θ_j is the latent ability level for examinee j, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and 1.7 is a scaling factor. The a_i parameter is analogous to item-total correlations in classical test theory (CTT; Spearman, 1987), and b_i is comparable to item proportion correct. The 2PL model can also be re-written using an intercept parameter $d_i = -1.7a_ib_i$. The 2PL model assumes monotonicity, unidimensionality, and local independence. Monotonicity is met when P_{ij} is a non-decreasing function of θ. The unidimensionality assumption requires that all test items measure only one latent trait, which is one of the core factors of a valid assessment (De Ayala & Hertzog, 1991; Haladyna & Rodriguez, 2013). The local independence assumption states that once θ is accounted for, item response patterns are uncorrelated to each other. Given these assumptions, the log likelihood of a response vector of n items for examinee j is

$$\sum_{i=1}^{n}[X_{ij}\ln P_{ij} + (1 - X_{ij})\ln(1 - P_{ij})] \tag{2}$$

The θ that maximizes this log likelihood is the maximum-likelihood estimate (MLE) of ability, represented as $\hat{\theta}$. Unlike in CTT where the sum score is the estimate of the true score, using θ allows invariance of ability, so multiple examinees taking a different set of items can be scored on the same score scale (Wu et al., 2016). A good item is one that has a high a_i and a b_i that matches the examinee’s θ, which will substantially lower the standard error of $\hat{\theta}$. Therefore, a healthy item bank will contain items with a variety of difficulty levels so the exams can accommodate both high and low performers. For more details on IRT, see Baker (2001) and Lord and Novick (2008). For applications of IRT in machine learning, see Martínez-Plumed et al. (2016).

Proposed AI Field-Testing Methodology

The proposed method requires a scenario with (a) a substantial number of previously field-tested and calibrated items accompanied by their item text, (b) a new set of field-test items that currently lack data aside from their text, and (c) these two groups of items are consistent in overall design variations and measurement constructs. The overall flow of AI field-testing is shown in Figure 1.

Process Item Text Data

I use English grammar multiple-choice items from a real large-scale assessment to illustrate the proposed approach. These items assess skills to apply or edit English grammar usage, capitalization, punctuation, and spelling. The stems of these items can vary greatly. For example, they may state, ‘‘Choose the sentence that is

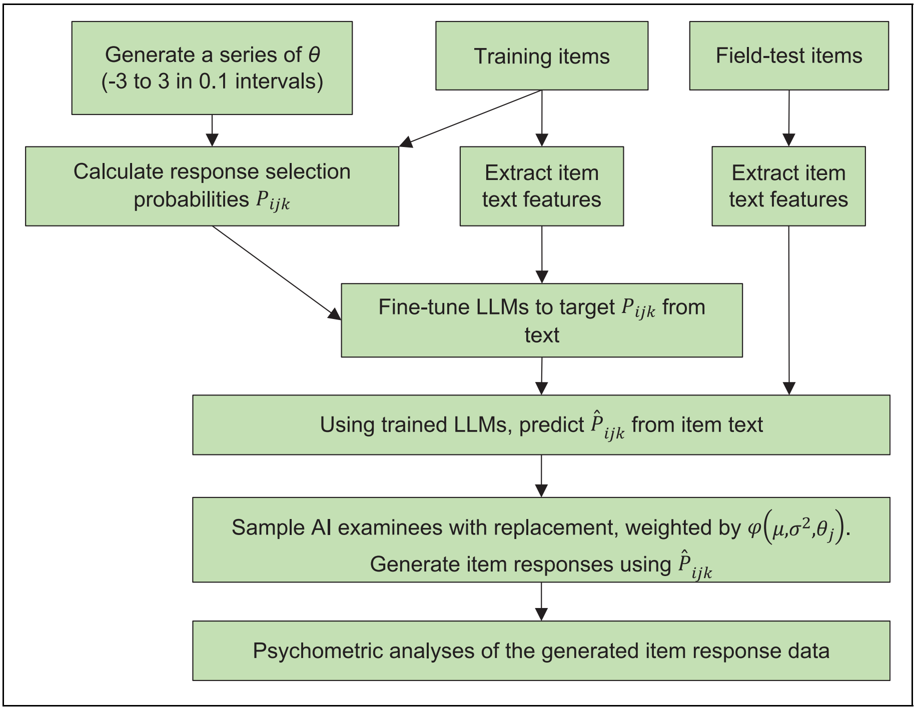
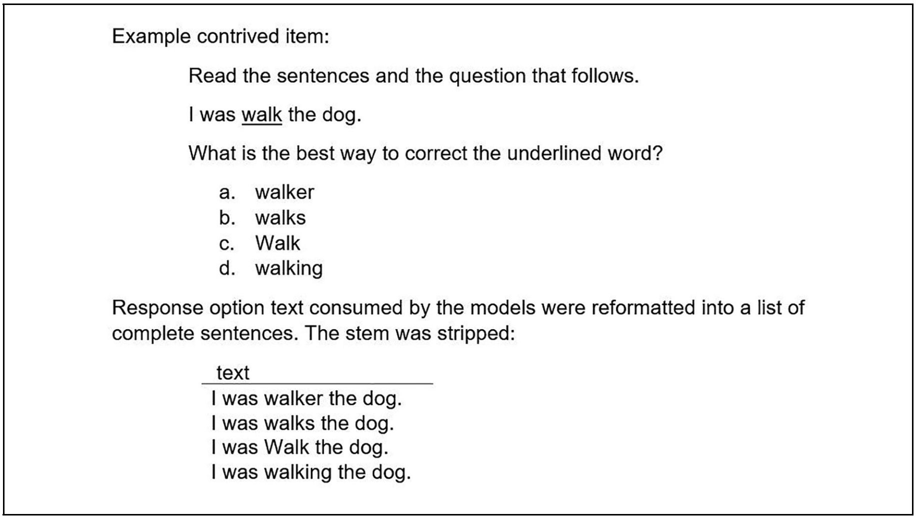

Generate a series of θ (-3 to 3 in 0.1 intervals)
Training items
Field-test items
Calculate response selection probabilities P_{ijk}
Extract item text features
Extract item text features
Fine-tune LLMs to target P_{ijk} from text
Using trained LLMs, predict $\hat{P}_{ijk}$ from item text
Sample AI examinees with replacement, weighted by $\varphi(\mu,\sigma^2,\theta_j)$.
Generate item responses using $\hat{P}_{ijk}$
Psychometric analyses of the generated item response data

Figure 1. Flowchart of AI Field-Testing.

Note. Using latent ability θ randomly generated for every LLM, the response selection probabilities of training items are calculated. Along with its item text, multiple LLMs are fine-tuned individually. Fine-tuned models are used to generate item responses using only the item text. Finally, item calibration and other psychometric analyses are conducted on the generated responses. AI = artificial intelligence; LLM = large-language model.



Example contrived item:

Read the sentences and the question that follows.

I was walk the dog.

What is the best way to correct the underlined word?

a. walker
b. walks
c. Walk
d. walking

Response option text consumed by the models were reformatted into a list of complete sentences. The stem was stripped:

text
I was walker the dog.
I was walks the dog.
I was Walk the dog.
I was walking the dog.

Figure 2. Example Restructuring of Multiple-Choice Question Text for LLM Consumption.
Note. LLMs were trained to identify whether sentences follow correct or incorrect English grammar. LLM = large-language model.

punctuated correctly'' or ''Choose the correct word to replace the underlined word in the sentence.'' The design inconsistencies in the stem may add unnecessary challenges for LLMs to understand the text. Therefore, the item text data are restructured. First, the item response option text is modified to be compatible with a stem that states ''Choose the sentence(s) with correct grammar usage'' (i.e., if not already compatible). The option text is reworded appropriately to one or more complete sentences with or without any grammatical errors (see Figure 2). Then, because now the stem is identical for all items, we can remove the stem from the input text data, and use the target label to guide the LLMs to find the sentences with correct grammar (see next section). Removing the stem also limits the data consumed by LLMs and speeds up the training. Note that to apply AI field-testing to other types of items, the item text processing procedure may need considerable adjustments.

Calculate Conditional IRT Probabilities

A total of 61 AI examinee ability levels θ_j of 3.0, 2.9, 2.8, . . . -3.0 are assigned to individual 61 LLMs, which span the majority of the target population ability distribution $\theta \sim N(\mu, \sigma^2)$, where $\mu = 0$ and $\sigma^2 = 1$. As LLMs take a long duration to fine-tune, limiting the number of models to 61 was a strategy to enhance the training efficiency. For every training item *i*, where the multiple-choice options are represented by *k*, calculate conditional probability of correct response option *c* based on the 2PL model: $P_{ijk=c} = P_{ij}$. Then, calculate the conditional probability of selecting each distractor (i.e., incorrect) option $k \neq c$. This can be based on marginal proportion of human examinees that selected each distracter response option $D_{ijk \neq c}$

$$P_{ijk \neq c} = \frac{D_{ijk \neq c}}{\sum_k D_{ijk \neq c}} (1 - P_{ijk=c}) \tag{3}$$

Therefore, the selection probabilities sum to $\sum_k P_{ijk} = 1$ within item i and examinee j. Although the nominal response model (Bock, 1972) could be used in this step to estimate the distractor selection probabilities with more precision, this is a rather complex and inconvenient model that most testing programs do not use to score examinees.

Fine-Tune Transformers With Item Options as "Separate_" Sequence Inputs

The item option text is tokenized and used to fine-tune the DeBERTa-v3-large transformer neural network (He et al., 2021) for a regression task, using P_{ijk} as the target label (i.e., the output). In extending the pre-trained language model, I augment the network by adding an extra fully connected layer, which serves as the new output layer. Adhering to the fine-tuning principles outlined in Devlin et al.'s (2018) study, I utilize the initial output of the pre-trained language model, corresponding to the special token ''[CLS].'' This is the sole output utilized for regression and classification tasks. The additional output layer consists of a single neuron, and the weights for connections with the preceding layer are randomly initialized. Throughout the fine-tuning process, both the weights of the additional layer and the internal weights of the pre-trained language model are updated.

A softmax function $\exp(l) / \sum \exp(l_k)$ is applied to the four output logits for each item, which ensures that the predicted selection probabilities sum to $\sum_k \hat{P}_{ijk} = 1$ within item i and model j. Since I am predicting probabilities, cross-entropy loss (CEL) is used to quantify and minimize the distance between the true and predicted probability distributions:

$$\text{CEL}_j = -\sum_{ik} \left[P_{ijk} \log\left(\hat{P}_{ijk}\right) + \left(1 - P_{ijk}\right) \log\left(1 - \hat{P}_{ijk}\right) \right] \tag{4}$$

This modeling approach is named the ''Separate_'' method.

Alternative Method: "Concatenate_" the Item Options

Instead of inputting the four option texts for each item separately, an alternative is to concatenate them into a single input sequence using a separator special token ''[SEP]'' between each option. The advantage may be that this gives the model the entire context of the item text for each response option prediction. We can still predict four individual selection probabilities by outputting four values per text sequence (i.e., item). However, one issue is that there is no direct way for the model to know which output value is associated with each option. A solution to this could be to input items in the training data with multiple duplications, each time reordering item option text and target probabilities. This way, the model may learn the connection between the item option text and probabilities based on their locations. As short hand, the model with no duplication in the training data is named ''Concatenate_1,'' and models with items repeated in the training data 4 and 24 times are named ''Concatenate_4'' and ''Concatenate_24,'' respectively. Concatenate_4 option text is ordered like 1234, 3412, 2143, and 4321, so each option is positioned at all four locations exactly once. Repeating each item 24 times with Concatenate_24 is the maximum possible ways that four options can be ordered ($4 \times 3 \times 2 \times 1 = 24$). Otherwise, the fine-tuning process is identical to the Separate_ method.

Generate Item Responses

For every training and field-test item *i*, use the fine-tuned LLM to obtain predicted probabilities $\hat{P}_{ijk}$. Then, from the pool of 61 LLMs, sample AI examinees with replacement weighted by $\varphi(\mu, \sigma^2, \theta_j)$, which is the normal density of θ. Finally, generate item responses based on $\hat{P}_{ijk}$. Note that I refer to the LLM as an ''AI examinee'' when it generates responses, because that is the point where the model can be treated like a human examinee. Although it may be possible to avoid generating any responses by directly using $\hat{P}_{ijk}$ for analyses such as item calibration, being able to use generated responses in existing field-testing workflows is practical and convenient.

Methods

Using real data, I simulate a scenario where I have a large number of previously calibrated items that can be used as training data, and wish to field-test a smaller set of new items that do not yet have any data, other than their text. Statistics previously obtained from real field-testing were treated as true values. Results from multiple modeling approaches were compared. Example Python code from this study is available on Github: https://github.com/hotakamaeda/ai_field_testing1

Assessment Data

The study included 466 items from a large-scale English language arts and literacy exam item bank. All items were four-option multiple-choice items used to assess 3rd to 8th and 11th grade students' skills to apply or edit English grammar usage, capitalization, punctuation, and spelling. These items have been previously field-tested among grade-appropriate students from the United States (mean student *N* per item = 3,161). Items have been calibrated on a vertical scale with the 2PL model.

Table 1. Number of Items and True Ability Distribution for Each Examinee Grade.

Grade level	Number of training items	Number of field-test items	μ (ability mean)	σ^2 (ability variance)
3	50	8	−0.88	0.61
4	50	7	−0.51	0.68
5	50	12	−0.14	0.70
6	48	9	0.06	0.67
7	57	7	0.26	0.74
8	49	11	0.44	0.73
11	92	16	0.78	0.95
All	396	70	0.00	1.00

Note. The ability mean and variance were obtained from past human examinee data using items included in the study. For easier interpretation, ability scaling has been adjusted so that the overall mean is 0 and variance is 1.

The items were randomly divided into about 85% training (n = 396) and 15% field-test items (n = 70). The training items are analogous to scored operational items that are calibrated and available in the item bank. They were used to train the LLMs, and also served as anchor items for calibration. The field-test items take the role of new items that are not available for training the LLMs. They were used to evaluate the performance of the trained models. See Table 1 for the number of items in each grade level. This table also shows μ_g and σ_g^2, which are the mean and variance of θ for grade g, respectively, assumed to be equal to their respective estimates derived from past human examinee data.

Separate_ and Concatenate_ Modeling Approaches

The DeBERTa-v3-large LLMs (He et al., 2021) were fine-tuned for the Separate_, Concatenate_1, Concatenate_4, and Concatenate_24 modeling approaches, each for all 61 ability levels θ_j of 3.0, 2.9, 2.8 . . . −3.0 (i.e., 4 × 61 = 244 models in total). To improve the efficiency of fine-tuning 61 models for each approach, I took advantage of the fact that, for example, the difference of a model with $\theta = 3$ and $\theta = 2.9$ was miniscule. Initially, the LLM was fine-tuned to have $\theta = 3$ with high epochs, then were later incrementally modified with low epochs while lowering θ by 0.1 each time. This technique substantially lowered the total number of epochs and training time needed to train all LLMs. AdamW optimizer (Loshchilov & Hutter, 2017) was used to minimize the CEL between the target and predicted probabilities (James et al., 2023). Fine-tuning was completed on a single NVIDIA A10G Tensor Core 24GB graphics processor, using the pytorch Python library (Paszke et al., 2019). Hyperparameters are shown in Table 2.

To create a representative item response data set for each grade level, LLMs were sampled *with replacement* 10,000 times weighted with probabilities $\varphi(\mu_g, \sigma_g^2, \theta_j)$.

Table 2. Model Training Data Size and Hyperparameters.

Model	Training step	Training data observation count	Learning rate	Weight decay	Batch size	Epochs
Separate_	Initial	396	4e-6	0.2	16	10
	Final	396	1e-6	0.2	16	2
Concatenate_1	Initial	396	7e-6	0.2	4	4
	Final	396	1e-6	0.2	4	2
Concatenate_4	Initial	1,584	7e-6	0.2	4	15
	Final	1,584	1e-6	0.2	4	2
Concatenate_24	Initial	9,504	7e-6	0.2	4	5
	Final	9,504	1e-6	0.2	4	1
Predicted_2PL	2PL *a*	396	1e-6	0.1	4	2
	2PL *d*	396	1e-6	0.1	4	2

Note. The "Initial" training step is when the model is initially trained to have a θ of 3 with a relatively high epoch. Then, this initial model was incrementally modified, lowering the θ by 0.1 each time using the hyperparameters listed in the "Final" training steps. 2PL = 2-parameter logistic model.

This represents the normal density of θ_j for mean and variance of θ for grade g (see Table 1). Probability resampling from 61 LLMs was more efficient than training 10,000 LLMs for every grade level. Although sample size of 1,000 is typically sufficient for estimating the 2PL model (Morizot et al., 2007), it was computationally cost-effective to raise the sample size to 10,000 for increased statistical precision. Item response data were generated based on $\hat{P}_{ijk}$ for all training and field-test items.

"Simulated_Human" Field-Testing Approach

Human field-testing was simulated as a point of comparison. Unlike AI, human examinees have a practical limit to the number of items they can respond to, which also limits the number of respondents per item.

Every field-test item was taken by 1,000 simulated human examinees with θ randomly generated from $N(\mu_g, \sigma_g^2)$, where μ and σ^2 were based on the grade level g of that field-test item (see Table 1). Item responses were generated directly from P_{ijk}. Simulated human examinees also responded to a random 40 training items associated with the respective grade level of the field-test item. In other words, this simulates a field-test setting where human examinees take grade-appropriate field-test items embedded within 40 other operational items.

IRT Calibration and CTT Analysis

For the Separate_, Concatenate_, and Simulated_Human approaches, field-test items were calibrated to the 2PL model one item at a time using the data set with the matching grade level. Mean and variance of θ were freely estimated. Available training items were anchored (i.e., fixed) on the true item parameters a_i and b_i so the field-test items are calibrated on the same scale as the training items. Calibration was completed using the mirt R package (Chalmers, 2012). In addition, proportion correct, item-total correlation (i.e., Pearson *r*), distractor selection proportion, and distractor discrimination were calculated for each field-test item (Haladyna & Rodriguez, 2013), again using the data set with the matching grade level.

"Predicted_2PL" a and d Parameters Modeling Approach

As an additional baseline comparison, the 2PL a and d parameters were predicted directly from the item text using DeBERTa-v3-large (He et al., 2021). This method is an attempt to replicate common difficulty prediction approaches found in the recent literature (AlKhuzaey et al., 2023; Benedetto, 2023; Benedetto et al., 2023). For each item, the four options were concatenated with the ''[SEP]'' separator special token, with the correct option text always in the first position. The mean squared error loss was minimized when fine-tuning the model. The 2PL b parameter was calculated based on the 2PL a and d by $b = -d/1.7a$. This baseline modeling approach was named ''Predicted_2PL.'' Compared to Separate_ and Concatenate_ approaches, Predicted_2PL uses the same training data, but is much simpler as it only requires two models (i.e., one for a and another for d) and skips the item response generation and calibration steps.

Latent Ability Estimation

To evaluate ability estimation accuracy, I generated a separate set of item response data to simulate a future scenario where field-test items have become operational scored items. A total of 10,000 simulated human examinees with ability sampled from $\theta \sim N(0, 1)$ responded to 40 random field-test items, where the items responses were generated from P_{ij}. Examinees were scored with MLE bounded within $[-4, 4]$ using the estimated 2PL parameters. Resulting estimates $\hat{\theta}$ were compared across all modeling approaches.

Evaluation

Statistics obtained from prior field-testing with real human examinees were treated as the true values. These were compared to the estimates obtained from Separate_, Concatenate_1, Concatenate_4, Concatenate_4, Predicted_2PL, and Simulated_Human approaches. Mean signed bias (i.e., mean of estimate minus true), root mean squared error (RMSE), and Pearson correlations (*r*) were calculated for every estimate.

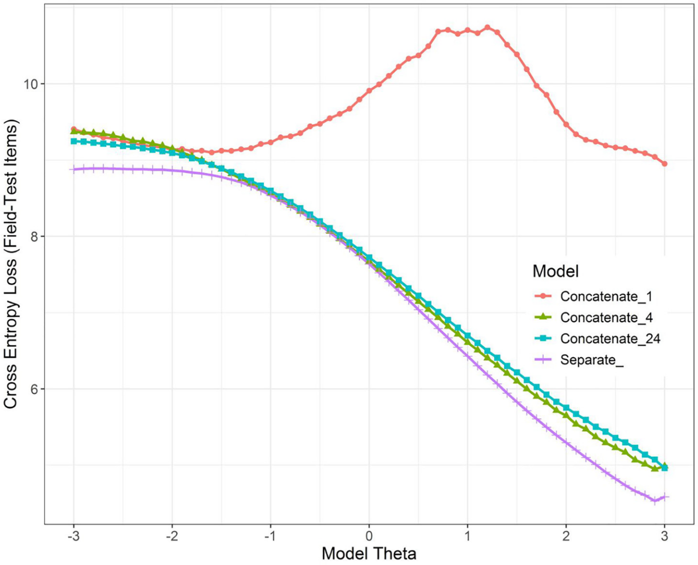


Figure 3. The Cross-Entropy Loss of the Field-Test Items, Based on the 61 Models for Each of the Four Modeling Approaches.

Note. The Separate_ method was clearly superior throughout the entire range of θ.

Results

Figure 3 shows the CEL values for the modeling approaches using the field-test items (evaluation data). Concatenate_1 clearly performed poorly, as it nearly consistently had the highest loss. Concatenate_4 and Concatenate_24 were nearly indistinguishable, despite Concatenate_24 requiring six times the amount of training data and time. The Separate_ method had the lowest CEL for all θ. Therefore, Concatenate_1 and Concatenate_24 models will be omitted from here on, and the focus will be on Separate_ and Concatenate_4.

The positive slopes in Figure 4 confirm that AI examinee θ influenced the probability of answering items correctly, as intended. However, slopes for the field-test items for Concatenate_4 and Separate_ were less steep than the training items and farther from the true P_{ijk}, which may indicate some overfitting and the limited generalizability of the trained models. Separate_ had a steeper slope that was closer to the truth compared to Concatenate_4.

Table 3 presents the summary comparisons of all selection probabilities, IRT parameters, and CTT statistics calculated among field-test items for all approaches. For both Concatenate_4 and Separate_, mean of $\hat{P}_{ijk}$ and mean of P_{ijk} were equivalent. Estimated statistics from Concatenate_4 and Separate_ all resulted in positive correlations with the true values. Separate_ approach was superior to Concatenate_4

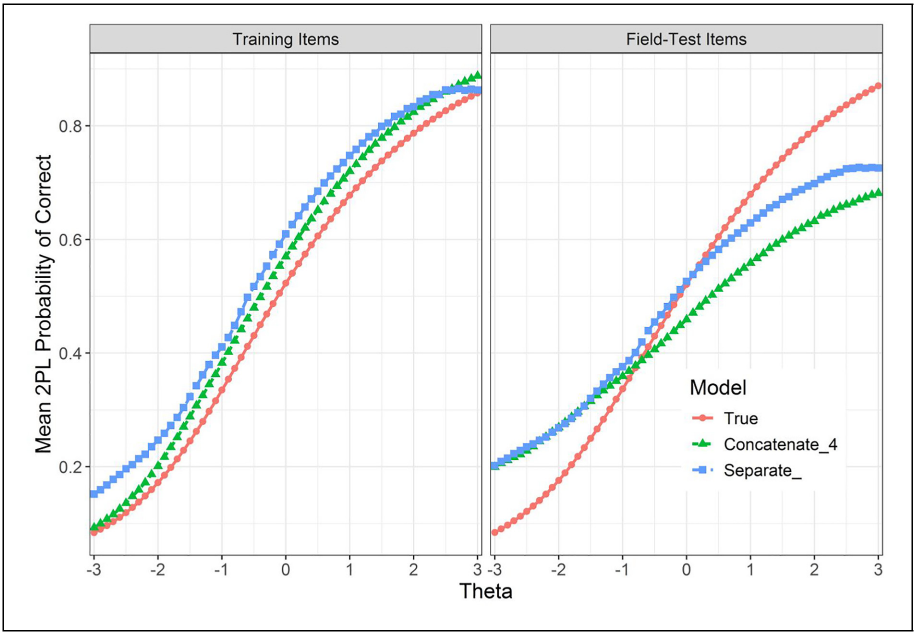


Figure 4. Difference in Slopes for Training and Field-Test Items for Concatenate_4 and Separate_ May Indicates the Limited Generalizability of the Trained Models.

from nearly all perspectives (i.e., lower bias and RMSE, higher correlation). However, compared to Simulated_Human, the correlations were consistently lower and RMSE was consistently higher for every statistic. An exception was for distractor discrimination, where Concatenate_4 (r = .29) and Separate_ (r = .33) had a higher correlation than Simulated_Human (r = .17). This may be a result of the study design, where the distractor correlations with θ were not simulated among the Simulated_Human group. Estimated statistics from Simulated_Human were all nearly unbiased. The Predicted_2PL baseline approach performed well in identifying the mean of the 2PL parameters, but the estimates had very narrow variances. For example, Predicted_2PL d had a SD of 0.09, even though the true SD was 1.17. This was sufficient, however, for a relatively accurate estimation of θ. But this method may not be useful for identifying poor quality items as it treats all items as having nearly identical parameters.

Separate_ Approach Item-Level Results

Figures 5 and 6 show the item-level results of Separate_, which reveal more details about its performance. A key limitation of Separate_ was that in 11 of the 70 field-test items, the 2PL a estimation failed catastrophically, and resulted in estimates of

Table 3. Estimation of IRT and CTT Statistics.

Statistic	Model	*M*	*SD*	Bias	RMSE	*r*
P_{ijk}	True	0.25	0.24			
	Concatenate_4	0.25	0.20	0.00	0.16	.77
	Separate_	0.25	0.21	0.00	0.14	.83
2PL P_{ij} ($P_{ijk=c}$)	True	0.50	0.31			
	Concatenate_4	0.45	0.24	0.04	0.21	.74
	Separate_	0.50	0.25	0.00	0.18	.82
2PL θ (ability)	True	0.00	1.01			
	Simulated_Human	0.01	1.08	0.01	0.41	.92
	Predicted_2PL	0.04	0.76	0.04	0.45	.91
	Concatenate_4	0.51	1.43	0.50	0.89	.87
	Separate_	0.26	1.25	0.26	0.62	.90
2PL *a* (discrimination)	True	0.55	0.25			
	Simulated_Human	0.56	0.25	0.01	0.06	.97
	Predicted_2PL	0.62	0.08	0.07	0.26	.04
	Concatenate_4	0.27	0.24	−0.28	0.36	.53
	Separate_	0.34	0.22	−0.21	0.31	.50
2PL *b* (difficulty)	True	0.16	1.19			
	Simulated_Human	0.13	1.18	−0.03	0.15	.99
	Predicted_2PL	0.00	0.08	−0.16	1.16	.42
	Concatenate_4	0.49	1.79	0.32	1.93	.22
	Separate_	0.22	1.42	0.06	1.76	.08
2PL *d* (intercept)	True	0.17	1.17			
	Simulated_Human	0.18	1.17	0.00	0.06	1.00
	Predicted_2PL	0.01	0.09	−0.17	0.26	.43
	Concatenate_4	−0.27	0.73	−0.44	0.36	.73
	Separate_	−0.13	0.64	−0.30	0.31	.71
Proportion correct	True	0.51	0.16			
	Simulated_Human	0.52	0.15	0.01	0.03	.99
	Concatenate_4	0.45	0.14	−0.06	0.17	.42
	Separate_	0.51	0.14	0.01	0.15	.49
Item-total correlation	True	0.31	0.10			
	Simulated_Human	0.22	0.10	−0.09	0.10	.93
	Concatenate_4	0.17	0.14	−0.15	0.20	.41
	Separate_	0.21	0.12	−0.11	0.16	.44
Distractor selection proportion	True	0.16	0.09			
	Simulated_Human	0.16	0.09	0.00	0.01	.99
	Concatenate_4	0.18	0.10	0.02	0.10	.46
	Separate_	0.16	0.09	0.00	0.10	.42
Distractor discrimination	True	−0.15	0.08			
	Simulated_Human	−0.18	0.07	−0.03	0.10	.17
	Concatenate_4	−0.09	0.08	0.06	0.12	.29
	Separate_	−0.12	0.08	0.04	0.10	.33

Note. The 2PL *b* parameter is bounded between [−3, 3] as some items had extreme values. 2PL = 2-parameter logistic model; RMSE = root mean squared error; IRT = item response theory; CTT = classical test theory.

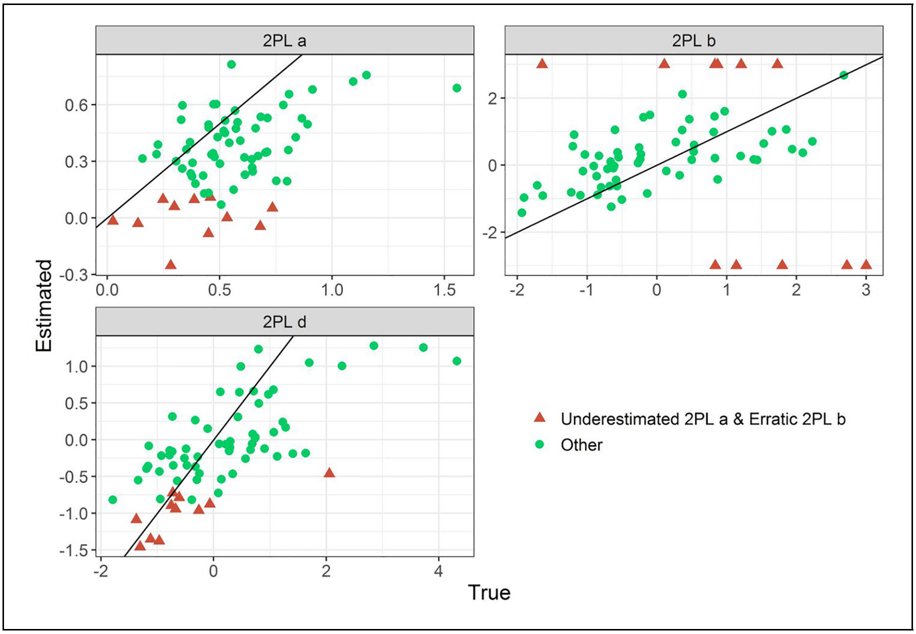


Figure 5. Separate_ Modeling Approach Can Fail Catastrophically for Some Items, Resulting in an Underestimated 2PL a and d, and Erratic b (2PL b Is Shown Bounded Within [−3, 3]).

near zero or negative, regardless of the true value (see Figure 5). This meant that, for example, two AI examinees with θ of 3 and −3 had similar probabilities of answering these items correctly. This resulted in an underestimated 2PL *a* and *d* and erratic 2PL *b*. For example, the 2PL *b* estimates bounded between [−3, 3] had bias = 0.06, RMSE = 1.76, and *r* = .08, but improved substantially without those 11 problematic items: bias = −0.23, RMSE = 0.92, and *r* = .59. Similarly, estimates of 2PL P_{ij} improve from bias = 0.00, RMSE = 0.18, and *r* = .82 to bias = −0.02, RMSE = 0.16, and *r* = .88. These problematic items are also visible in Figure 6, where $\hat{P}_{ij}$ was nearly independent of θ and P_{ij}.

Dimensionality

Unidimensionality is fundamental to test validity (Nunnally & Bernstein, 1994). Dimensionality of the AI examinee item response data was examined. Given that the target probabilities used to fine-tune the LLMs were generated from a unidimensional 2PL model, we can expect the generated responses to be unidimensional as well. Field-test item response data from a randomly generated 5,000 AI examinees from all grade levels were used. The first four eigenvalues for both the Separate_ (5.34, 0.28, 0.21, and 0.19) and Concatenate_4 (4.44, 0.23, 0.21, and 0.21) showed a sharp drop

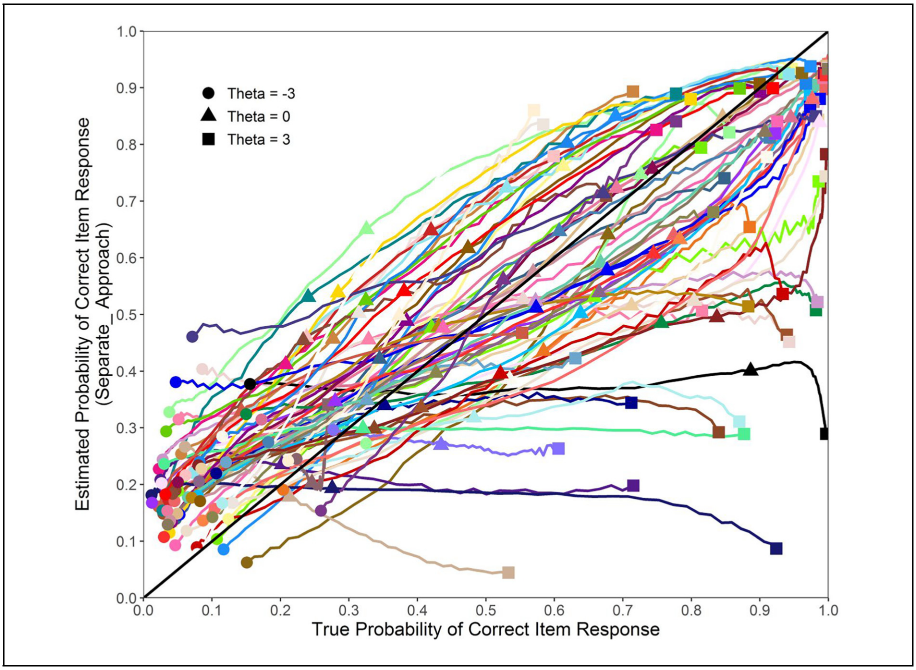


Figure 6. Separate_ Modeling Approach Used to Estimate Field-Test Item 2PL P_{ij} (Conditional Probability of Correct Response).

Note. Every line and color represent one of 70 field-test items. Each line shows 61 models strung together with θ ranging from 3 to −3. Each line begins with θ = −3 and ends with θ = 3. Perfect model fit would show a straight 45° positive slope for every item. Lines with nearly zero or negative slope are items with catastrophic modeling failure, where $\hat{P}_{ij}$ was nearly independent of θ and P_{ij}.

after the first value, which strongly indicated that these set of items were unidimensional (Cattell, 1966). Using the lavaan R package (Rosseel, 2012), confirmatory factor analysis (CFA) was conducted to confirm that the test was unidimensional (Bandalos & Finney, 2018). The estimator = ''WLSMV'' option was used as recommended for ordered categorical variables (Flora & Curran, 2004; Rhemtulla et al., 2012). Based on conventional standards (Hu & Bentler, 1999), fit statistics all indicated nearly perfect fit to the unidimensional CFA model for the Separate_ approach, $\chi^2(2{,}345) = 2{,}301.2$, $p = .74$, Tucker-Lewis index (TLI) = 1.001, CFI = 1.000, RMSE = .000, standardized root mean square residual (SRMR) = .021. Concatenate_4 had a marginally worse fit, but still very good, $\chi^2(2{,}345) = 2{,}441.59$, $p = .080$, TLI = .997, CFI = .997, RMSE = .003, SRMR = .023. Standardized factor loadings for Separate_ ranged from −0.16 to 0.59 ($M = 0.31$, $SD = 0.18$). Standardized factor loadings for Concatenate_4 ranged from −0.11 to 0.74 ($M = 0.26$, $SD = 0.20$).

Discussion

The ultimate goal of AI field-testing is to enhance the efficiency of field-testing by completely replacing human examinees with AI counterparts, and generate item response data that can be used for various statistical evaluations of item quality that are essential in traditional field-testing. The proposed approaches integrated LLMs and IRT to train AI to mimic human responses to English grammar multiple-choice items. The current study demonstrated that IRT- and CTT-based statistics calculated using AI item responses showed moderate resemblance of those obtained from human examinees. Latent ability estimation based on item parameter estimates from AI field-testing was similar to ability estimates based on the ‘‘true’’ item parameters that were originally used to fine-tune the AI models. AI-generated responses may also have the potential to be used for dimensionality analyses, which is typically plagued with data sparseness issues in computerized adaptive tests (Bulut & Kim, 2021).

Unlike AI, human examinees are limited in the number of items they can complete. This also lowers the sample size for each field-test item. These two factors limit the quality of human examinee item response data. Regardless, the study showed that human field-testing is still superior to AI field-testing for accurately calibrating items or calculating item statistics. One reason may be because the item text had to be reformatted for LLM consumption, including a complete elimination of the stem text. This was necessary as these items were not originally designed to be processed by LLMs. Past studies show that simple changes to the item text can change the item difficulty noticeably (Byrd & Srivastava, 2022). Especially considering AI examinees cannot think or feel like humans do, there may always be an inherent gap between how AI and humans respond to exam items.

Among the AI field-testing methods included in the article, the Separate_ method performed far more accurately than Concatenate_1, Concatenate_4, and Concatenate_24. Furthermore, Separate_ is more convenient than Concatenate_ approaches as it takes less training time, can handle longer input sequence text, and can elegantly accommodate items with varying numbers of response options.

Importantly, the proposed approaches preserved the core structure of traditional field-testing by replacing human examinees with AI counterparts, allowing most analyses and calibration procedures to proceed as usual. This seamless integration into existing workflows not only offers convenience and practicality but also makes the technology more accessible, which may help stakeholders better understand and accept it. Building this understanding and trust is crucial for the successful adoption of new technologies like AI in education (Aloisi, 2023; Nazaretsky et al., 2022; Qin et al., 2020).

Trust in AI may also be closely tied to the widespread concerns about AI bias and ethics (Bolukbasi et al., 2016; Caliskan et al., 2017; Hassija et al., 2023; Morales et al., 2023). Before AI field-testing can be confidently applied in high-stakes exams, it is essential to develop and implement robust methods to mitigate AI bias. In psychometrics, item bias against examinees based on their background is typically identified through differential item functioning (DIF) analysis (Holland & Wainer, 1993).

As DIF was not addressed in the current article, research is needed to integrate it into the AI field-testing framework.

Another key issue was how AI examinees' response behavior to the training items and field-test items was fundamentally different (see Figure 4). This mattered as training items were used as anchors during calibration. This likely had a negative effect on all 2PL parameter estimations. A solution could be to make an initial mutually exclusive distinction of three item groups: (a) calibrated training items, (b) calibrated anchors, and (c) new field-test items. This approach to avoid using training items as anchors may eliminate some issues. Another option may be to fix the AI θ to its true values, which eliminates the need for anchor items. Other alternatives could be to raise training item sample size, or adjust training hyperparameters so that the model is more generalizable to the field-test items.

While this article primarily presented a simulation study, simulated data were used sparingly. Simulated data were employed to create the Simulated_Human group and to assess the accuracy of latent ability estimation across all modeling approaches. However, the item parameters and statistics, the mean and standard deviation of θ for both the overall group and individual grades, and the item text were all derived directly from an ongoing large-scale assessment program. Notably, the data used for fine-tuning the LLMs consisted entirely of real data. In other words, the study was conducted on real AI examinees, rather than simulated ones.

The current article used relatively simple English grammar multiple-choice items to illustrate and evaluate AI field-testing. Answering these items required only the knowledge of the English language, which made them exceptionally appropriate for testing English-based LLMs. In the literature, evidence shows that LLMs can be trained to answer mathematical problems (Xu et al., 2024) or medical items that require knowledge and reasoning (Singhal et al., 2023). Whether the proposed techniques can generalize to these other item types will need further context-specific investigation.

The literature provides examples of using NLP to handle varied item types. For instance, Settles et al. (2020) employed machine learning algorithms to generate an entire English language assessment. They first determined vocabulary difficulty using word frequency, word character lengths, and expert judgment. Next, they ranked passage difficulty based on average word and sentence length, word frequency, pre-labeled online passages, and expert evaluations. The predicted vocabulary and passage difficulty were then used to generate five item formats assessing listening, speaking, reading, and writing skills. Their use of multiple techniques highlights key considerations for predicting the difficulty of items with varied formats.

Benedetto (2023) may have reported the most comprehensive quantitative comparison of modern item difficulty prediction methods. The author compared BERT (Devlin et al., 2018) and DistilBERT (Sanh et al., 2019) transformers with random forest regression models that utilize linguistic features, readability indices, term frequency—inverse document frequency (Manning et al., 2008), and word2vec embeddings (Mikolov et al., 2013). Transformers were the most predictive of item difficulty. However, the accuracy depended heavily on the exam type, where R^2 ranged from .19 to .62 with LLMs trained on 4,000 to over 100,000 items. Increasing the number of training items improved the R^2. The best performing BERT models from Benedetto (2023) were partially replicated in the current study with the Predicted_2PL approach. Predicting the 2PL b parameter with Predicted_2PL had an R^2 of .18. This relatively poor performance is likely due to the low item sample size and the particular type of items used in the current article. The results of the current study may improve drastically if models are trained on more data.

Limitations

The proposed method uses item text to predict IRT selection probabilities with regression. An alternative would be to predict the response selection categories instead, which was Lalor et al.'s (2019) approach. Predicting responses was substantially less convenient and less efficient than 2PL probabilities. However, if the ultimate goal is to mimic all aspects of human response behavior, training LLMs directly based on raw human examinee response data could be a necessary future direction, as many nuanced patterns are not captured by the 2PL model. This is evident in the extremely good fit of the unidimensional CFA to the AI examinee response data, suggesting that minor multidimensionality was lost in the process.

Similarly, I did not directly compare AI and human examinee item response patterns in this article. Rather, I compared item-level statistics obtained from such responses, such as proportion correct or the item parameters. Although these comparisons were sufficient in showing the model performance, comparing responses directly could be the ideal approach.

Furthermore, the parameters and statistics obtained from prior field-testing with real humans were treated as true values. In reality, these were estimates as well. This limitation likely inflated the accuracy of the Simulated_Human approach. The current article used English grammar multiple-choice questions to demonstrate the effectiveness of AI field-testing. However, English literacy exams are often not limited to these types of items or content. Research is necessary to develop LLMs that can respond to various types of items, or methods to integrate multiple LLMs that are designed to respond to specific item types. Incorporating additional text features in the LLM (e.g., item word count) may be another route to enhancing the approach, which may further close the distance between AI and human item response behavior.

Conclusion

This article presented an innovative approach of replacing human examinees with AI examinees for field-testing newly written exam items. The study demonstrated that AI item response data can be used to calculate item statistics and conduct item calibration, distractor analysis, dimensionality analysis, and latent trait scoring. Although AI field-testing still fell short of the accuracy of item calibration and analyses that were performed with human examinee response data, the potential resource savings in transitioning from human to AI field-testing cannot be understated. AI could shorten the field-testing timeline, prevent human examinees from seeing low-quality field-test items in real exams, shorten test lengths, eliminate item exposure, test security, and sample size concerns, and reduce the overall cost. In the era of generative AI, the research on automatic item generation may tend to far outpace the rate at which items can be field-tested, which could result in a bottleneck. A strategic combination of automatic item generation and AI field-testing may enable an extremely efficient expansion of the item bank. Researchers are encouraged to explore methods to enhance AI examinees to be more reflective of human examinee behavior, as well as generalize its capabilities to handle various types of items.

Author's Note

Preprint is available here: Hotaka Maeda. Field-testing multiple-choice questions with AI examinees: English grammar items, 29 July 2024, PREPRINT (Version 2) available at Research Square [https://doi.org/10.21203/rs.3.rs-3858355/v2]

Acknowledgments

The author appreciates the support and feedback of Aren Simmons, Catherine Francis, Fang Peng, Rochelle Michel, Shumin Jing, and Yikai ''EK'' Lu.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Hotaka Maeda https://orcid.org/0009-0000-9498-786X

References

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2023). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 1–53. https://doi.org/10.1007/s40593-023-00362-1

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, *58*(1), 98–110.

Baker, F. B. (2001). *The basics of item response theory*. https://eric.ed.gov/?id=ED458219

Bandalos, D. L., & Finney, S. J. (2018). Factor analysis: Exploratory and confirmatory. In Hancock, G. R., Stapleton, L. M., & Mueller, R. O. *(Eds.), The reviewer's guide to quantitative methods in the social sciences* (pp. 98–122). Routledge.

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517–530.

Benedetto, L. (2023). *A quantitative study of NLP approaches to question difficulty estimation* (arXiv preprint arXiv:2305.10236). https://arxiv.org/abs/2305.10236

Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., & Turrin, R. (2021). On the application of transformers for estimating the difficulty of multiple-choice questions from text. In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis & T. Zesch (Eds.), *Proceedings of the 16th workshop on innovative use of NLP for building educational applications* (pp. 147–157). Association for Computational Linguistics.

Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. (2020a). R2DE: A NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the 10th international conference on learning analytics & knowledge* (pp. 412–421). Association for Computing Machinery.

Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. (2020b). Introducing a framework to assess newly created questions with natural language processing. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin & E. Millán (Eds.), *International conference on artificial intelligence in education* (pp. 43–54). Springer.

Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, *55*(9), 1–37.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29. https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bulut, O., & Kim, D. (2021). The use of data imputation when investigating dimensionality in sparse data from computerized adaptive tests. *Journal of Applied Testing Technology*, *22*(2), Article 158509.

Byrd, M., & Srivastava, S. (2022). Predicting difficulty and discrimination of natural language questions. In S. Muresan, P. Nakov & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics: Short papers 602* (*Vol. 2*, pp. 119–130). Association for Computational Linguistics.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Choi, I. C., & Moon, Y. (2020). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, *17*(1), 18–42.

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*, Article 858273.

Conejo, R., Guzmán, E., Perez-De-La-Cruz, J.-L., & Barros, B. (2014). An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications*, *41*(2), 594–606.

De Ayala, R. J., & Hertzog, M. A. (1991). The assessment of dimensionality for use in item response theory. *Multivariate Behavioral Research*, *26*(4), 765–792.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* (arxiv preprint arxiv:1810.04805). https://arxiv.org/abs/1810.04805

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082–1116.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Taylor & Francis.

Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., & Sikdar, B. (2023). Unleashing the potential of conversational AI: Amplifying chat-GPT's capabilities and tackling technical hurdles. *IEEE Access*, *11*, 143657–143682.

He, P., Gao, J., & Chen, W. (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing* (arxiv preprint arxiv:2111.09543). https://arxiv.org/abs/2111.09543

He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with disentangled attention* (arxiv preprint arxiv:2006.03654). https://arxiv.org/abs/2006.03654

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Routledge.

Hsu, F. Y., Lee, H. M., Chang, T. H., & Sung, Y. T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, *54*(6), 969–984.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer.

Jiao, H., & Lissitz, R. W. (Eds.). (2020). *Application of artificial intelligence to assessment*. Information Age.

Lalor, J. P., Wu, H., & Yu, H. (2019). Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the conference on empirical methods in natural language processing: Conference on empirical methods in natural language processing* (*Vol. 2019*, p. 4240). NIH Public Access.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* (arXiv preprint arXiv:1907.11692). https://arxiv.org/abs/1907.11692

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge. https://doi.org/10.4324/9780203056615

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization* (arxiv preprint arxiv:1711.05101). https://arxiv.org/abs/1711.05101

Loukina, A., Yoon, S. Y., Sakano, J., Wei, Y., & Sheehan, K. (2016, December). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3245–3253). The COLING 2016 Organizing Committee.

Maeda, H. (2023). *Field-testing items using artificial intelligence: Natural language processing with transformers* (arxiv preprint arxiv:2310.11655). https://arxiv.org/abs/2310.11655

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2016). Making sense of item response theory in machine learning. In *ECAI2016* (pp. 1140–1148). IOS Press. https://dl.acm.org/doi/pdf/10.3233/978-1-61499-672-9-1140

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26. https://arxiv.org/abs/1310.4546

Mislevy, R. J. (1988). Exploiting collateral information in the estimation of item parameters. *ETS Research Report Series*, *1988*(2), i–31.

Morales, S., Clarisó, R., & Cabot, J. (2023). Automating bias testing of LLMs. In *2023 38th IEEE/ACM international conference on automated software engineering (ASE)* (pp. 1705–1707). Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/10298519

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). Guilford Press.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, *53*(4), 914–931.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*. https://arxiv.org/abs/1912.01703

Qin, F., Li, K., & Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British Journal of Educational Technology*, *51*(5), 1693–1710.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., & Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? In C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing* (pp. 4486–4503). Association for Computational Linguistics.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arxiv preprint arxiv:1910.01108). https://arxiv.org/abs/1910.01108

Settles, B. T., LaFlair, G., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180.

Spearman, C. (1987). The proof and measurement of association between two things. *The American Journal of Psychology*, *100*(3/4), 441–471.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. https://arxiv.org/abs/1706.03762

Wang, S., & Jiao, H. (2011). Incorporating person covariates and response times as collateral information to improve person and item parameter estimations. *Online Submission*. https://files.eric.ed.gov/fulltext/ED522409.pdf

Wu, M., Tam, H. P., & Jen, T. H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer.

Xu, Y., Liu, X., Liu, X., Hou, Z., Li, Y., Zhang, X., & Dong, Y. (2024). *ChatGLM-Math: Improving math problem-solving in large language models with a self-critique pipeline* (arXiv preprint arXiv:2404.02893). https://arxiv.org/abs/2404.02893

Yaneva, V., Baldwin, P., & Mee, J. (2019). Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán & T. Zesch (Eds.), *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 11–20). Association for Computational Linguistics.

Zhou, Y., & Tao, C. (2020). Multi-task BERT for problem difficulty prediction. In *2020 international conference on communications, information system and computer engineering (CISCE)* (pp. 213–216). Institute of Electrical and Electronics Engineers.