Using item features to calibrate educational test items: Comparing artificial intelligence and classical approaches

American Journal of Education and Learning
Vol. 10, No. 2, 178-189, 2025
e-ISSN:2518-6647

Peter Tran[1]
Wanyong Feng[2]
Stephen G. Sireci[3]
Hunter McNichols[4]
Andrew Lan[5]

(Corresponding Author)

[1,2,3,4,5]*University of Massachusetts Amherst, Massachusetts, USA.*
[1]*Email: wanyongfeng@umass.edu*
[2]*Email: andrewlan@cs.umass.edu*
[3]*Email: sireci@acad.umass.edu*
[4]*Email: petertran@umass.edu*
[5]*Email: wmcnichols@umass.edu*

ABSTRACT

Educational test items are typically calibrated onto a score scale using item response theory (IRT). This approach requires administering the items to hundreds of test takers to characterize their difficulty. For educational tests designed for criterion-referenced purposes, characterizing item difficulty in this way presents two problems: one theoretical, the other practical. Theoretically, tests designed to provide criterion-referenced information should report test takers' performance with respect to the knowledge and skills they have mastered, rather than on how well they performed relative to others. The traditional IRT calibration approach expresses item difficulty on a scale determined solely by test takers' performance on the items. Practically, the traditional IRT approach requires large numbers of test takers, who are not always available and who are not always motivated to do well. In this study, we use the construct-relevant features of test items to characterize their difficulty. In one approach, we code the item features; two other approaches are based on artificial intelligence (chain-of-thought prompting and LLM finetuning). The results indicate the coding and LLM finetuning approaches reflect the difficulty parameters calibrated using IRT, accounting for approximately 60% of the variation. These results suggest educational test items can be calibrated using construct-relevant features of the items, rather than only administering them to samples of test takers. Implications for future research and practice in this area are discussed.

Keywords: *Artificial intelligence, Educational testing, Item response theory, Large language models, Test development, Validity.*

DOI: 10.55284/ajel.v10i2.1543
Citation | Tran, P., Feng, W., Sireci, S. G., McNichols, H., & Lan, A. (2025). Using item features to calibrate educational test items: Comparing artificial intelligence and classical approaches. *American Journal of Education and Learning, 10*(2), 178–189.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
History: Received: 9 April 2025/ Revised: 10 June 2025/ Accepted: 22 July 2025/ Published: 29 August 2025
Publisher: Online Science Publishing

Highlights of this paper:

- For over 100 years, educational test items have been evaluated and calibrated through item analysis based on students' responses to items.
- However, by coding relevant content attributes of the items, estimates of item difficulty can be obtained using linear regression.
- Large language models can also capture item difficulty, but do not require coding attributes or large numbers of students; thus, they are likely to be the preferred method for calibrating test items in the future.

1. INTRODUCTION

Traditional and current practices in educational testing involve reporting scores on a scale derived from test takers' responses to test items. Classical approaches for placing test items and test takers on the same scale began with the work of Thurstone (1925) who used the proportions of test takers who correctly answered an item, and the ages of the students who responded to it, to account for the variability of the difficulty of the items. Beginning in the 1950s, this idea was expanded and improved using *item response theory* (IRT) (Lord, 1952; Lord & Novick, 1968; Rasch, 1960) which posits specific mathematical models representing the probability of a test taker earning a specific score on a test item, conditional on the difficulty of the item, the proficiency of the test taker, and in some cases, additional parameters associated with the item (discrimination, lower asymptote, etc.) (Hambleton, Swaminathan, & Rogers, 1991; Thissen & Steinberg, 2020)).

Although IRT is the dominant approach for creating score scales in educational testing, it has several shortcomings for contemporary educational testing practices. A first shortcoming is that the approach is inconsistent with the logic of *criterion-referenced testing*, which strives to provide information about how well a student has mastered a specific domain of knowledge and skills. For example, teachers may want to know whether students have fully grasped the notion of ratios and proportions before teaching students algebra. IRT and other traditional approaches for scaling educational tests define the score scale using either features of the population of test takers (e.g., mean test score) or the average difficulty of the items, which is also derived from test takers' performance. Thus, these scaling procedures are essentially *norm-referenced*, in that the interpretation of the scale and performance on it is relative to some norm group. In norm-referenced testing, test takers' performance on a test is interpreted relative to the performance of other test takers (not to the targeted domain of knowledge and skills). Scale scores, percentile ranks, grade equivalent scores, and many other score reporting metrics are derived from this norm-referenced perspective. The fact that virtually all educational tests use a norm-referenced scaling procedure to report students' performance on criterion-referenced tests has been largely overlooked and limits the validity of contemporary educational tests (Sireci, 2021).

A second significant limitation is the substantial data collection required for calibrating items using IRT. Items must be administered to at least 200 test takers for even the simplest IRT models (Hambleton et al., 1991) and to build an item bank of sufficient size, it typically requires pilot-testing the items on thousands of test takers. There was a time when doing so was only somewhat arduous, but as the COVID-19 pandemic illustrated, test takers are not always available for pilot testing. A related problem is that test takers are often not motivated to try their best on educational tests, particularly in the context of low-stakes testing, which leads to inaccurate IRT item difficulty estimates (Wise, Im, & Lee, 2021).

The purpose of the current study is to propose and explore methods for calibrating items onto a score scale using the construct-relevant features of the items to account for the variation in their difficulty, or, more appropriately, their complexity. If we can scale items based on their features relevant to the knowledge and skill domains measured, we will improve the validity of the interpretations made on the basis of test scores, as scores on

this scale will be explicitly associated with the complexity features of the items to which test takers responded. Additionally, the data required for calibrating the items will not depend on samples of test takers that may be unmotivated, unrepresentative, or unavailable.

As this is an initial foray into this area, our study involves investigating the item features on a math test and using approaches based on manual coding (suggested by literature from the 1980s) as well as newer approaches based on artificial intelligence (AI). Before presenting our methodology and results, we first provide a brief review of prior related research.

1.1. Prior Related Research

Research on simplifying or improving IRT item calibration has included both methods for calibrating items using smaller samples and coding items to improve test-based score interpretations. With respect to improving traditional item calibrations using smaller sample sizes, Swaminathan, Hambleton, Sireci, Xing, and Rizavi (2003) used subject matter experts' judgments of item difficulty as priors in IRT item parameter estimation and found substantial improvements in the estimation of IRT difficulty with sample sizes as small as 100. They concluded that such judgmental information could be used to reduce sample size requirements by 50%.

More germane to the purpose of the present study is research focused on deriving "construct" meaning from the underlying IRT scale to facilitate the interpretation of students' proficiencies concerning the domain measured. Fischer (1973) introduced the linear logistic latent trait model to incorporate complexity factors to represent the IRT item difficulties. Similarly, Whitely (1983) proposed multicomponent latent trait modeling to include a task decomposition model of item difficulties into a Rasch IRT model. Sheehan and Mislevy (1990) expanded on this research to relate IRT item parameters to relevant item features in a cognitive model based on extensions of Fischer's linear logistic test model. Using these features, they were able to predict approximately 80% of the IRT item difficulty parameters for items from the NAEP document literacy scale.

The research of Fischer (1973) and Whitely (1983) and Sheehan and Mislevy (1990) focused on explaining or predicting the item difficulty parameters from an IRT model. Although the current study also uses a predictive context (i.e., similarity of our complexity values to pre-calibrated IRT difficulty parameters), our ultimate goal is to replace the IRT difficulty scale derived from students' responses to items with a difficulty scale derived from the complexities of the content of the items themselves. This approach would transform the calibration and scaling of educational test items from norm-referenced approaches to criterion-referenced approaches based on the nature of the difficulty of an item with respect to the construct it measures. In the context of criterion-referenced testing, this transformation will better match the scaling model with the testing purpose and lead to a more valid assessment.

There are two main approaches for item calibration that are solely based on the items' textual information: expert-driven and machine-driven approaches. The expert-driven approach relies on educational experts' judgment, based on their domain knowledge and teaching experience, to calibrate items (Ling, Kang, Johns, Walls, & Bindoff, 2008). However, this approach is both time-consuming, subjective, and hard to scale to a large number of items (Benedetto, Cappelli, Turrin, & Cremonesi, 2020).

A machine-driven approach uses the item text and ground-truth difficulty value to extract different types of features from the item text. The extracted features are used to train the model. For example, Yaneva, Baldwin, and Mee (2020) used a random forest method to minimize the difference between predicted and ground-truth difficulty values. This task can be treated as either a regression or classification task, depending on whether the ground-truth value is continuous or discrete. There are two main kinds of features in this approach: syntax-level and semantic-level. Syntax-level focuses on the surface-level features of the item text. For example, the Flesch-Kincaid readability

score measures how difficult it is to understand the item text based on the average length of sentences and words, where larger values imply greater difficulty in understanding (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975). Semantic-level focus on the meaning of the item text; for example, using Word2Vec or long short-term memory (LSTM) deep neural networks to convert the item text into vectors that contain the semantic information of the item text (Cheng et al., 2019; Yaneva, Baldwin, & Mee, 2019). Recently, applications of large language modeling (LLM) have indicated success in understanding natural language and solving language-related problems. For example, Benedetto et al. (2021) used the BERT model to extract semantic features (see also AlKhuzaey, Grasso, Payne, and Tamma (2024) for a more detailed overview for the Machine-driven approach).

In summary, prior extensions of IRT modeling have illustrated that item features can be used to reflect item complexity and enhance understanding of students' test performance. More modern features based on LLMs illustrate that they can extract item features to quantify item complexity. Thus, comparing these approaches seems warranted.

2. METHOD

The purpose of this study is to determine whether content features and other attributes of mathematics items can be used to derive item "difficulty" parameters, such as those used in IRT. A related purpose is to explore three statistical methods for deriving item difficulty from item features, two of which are based on AI. Our specific research questions are.

a) Can the attributes of math items be used to create an item difficulty scale, similar to that used in IRT?

b) How well do item features predict difficulty parameters of mathematics items?

c) Which methods best predict item difficulty from item attributes: linear regression, chain-of-thought (CoT) prompting, or fine-tuning an LLM?

With respect to question (a), although we do not explicitly create an item difficulty scale, the results of our procedure could be used for this purpose. That is, we focus on predicting or recapturing the current IRT difficulties, but the predicted values could be used as actual scale values.

2.1. Data and Preprocessing

We used data from the Massachusetts Adult Proficiency Test (MAPT) for Math (Zenisky et al., 2024). Our data consisted of four-option multiple-choice math items. Additionally, we were provided with metadata that aided in feature selection: Flesch-Kincaid readability index, MAPT Level (curriculum level to which the item was written), math domain group, and cognitive level. The MAPT Level ranged from Level 2 to 5, where higher levels signified a higher grade level targeted by the item writer. There are 7 domain groups measured on the MAPT for Math: Numbers and Operations - Base 10, Operations and Algebraic Thinking - Expression and Equations, Geometry, Measurement and Data / Statistics and Probability, Numbers and Operations - Fractions / Ratios and Proportional Relationships, and Functions. Finally, there were 3 cognitive domains: Procedural Understanding (PU), Conceptual Understanding (CU), and Strategic Thinking (ST).

The native format of the input data was in XML. We extracted 517 active items from the XML file and stored them in JSON and XLSX formats. We performed automated cleaning by removing lingering XML syntax from the item stem and options. Next, we manually reviewed items to ensure the data reflected the essence of the actual items and to verify that the items provided all the information needed for the test taker to find the correct solution. We noted 191 items that contained images and saved them for a future study that incorporates images in model building. In total, we had 322 clean items as input for our models. The known (operational) IRT difficulty

parameters were calibrated using 3PL (see (Zenisky et al., 2024)) and served as the "true" values for our evaluation criteria.

2.2. Data Analysis

We used three methods to estimate the difficulties (complexity) of these math items. One approach was based on coding item features and using multiple linear regression; the other two were based on AI.

Method 1: Linear regression

The linear regression approach did not restrict the range of the predicted difficulty values. We used 13 construct-relevant features of items as input (predictors) to the model. Six of these predictors were linguistic features (number of sentences, number of words, number of nouns, number of unique nouns, Flesch-Kincaid score, and number of propositions). The seven other features were mathematical: number of numerical values, number of text numerical values, number of operators, number of unique operators, MAPT Level (analogous to the grade level to which the item was written), math domain group, and cognitive level. There were five MAPT Levels, seven domain groups, and three cognitive levels (see (Zenisky et al., 2024)).

Method 2: Chain-of-thought prompting

The second method used was Chain-of-Thought (CoT) prompting (Wei et al., 2022) with GPT-4o (Hurst et al., 2024). We restricted the range of the prediction to be between the minimum and maximum b values found in the dataset. We asked GPT-4o to generate reasoning steps to analyze the math item text across three aspects before predicting the difficulty value:

1) Mathematical Concepts Involved: describe the specific mathematical principles and theories that are required to understand and solve this question

2) Computational Complexity and Student Familiarity: assess the difficulty level of the calculations required to solve the question and consider the typical student's familiarity with these types of calculations and computational methods at their current education level.

3) Clarity and Potential for Misleading or Tricky Wording: evaluate the clarity of the question's wording and answer choices, and discuss any elements that might confuse test-takers or lead them towards incorrect interpretations or answers.

The method is referred to as zero-shot because the model is not provided with in-context training on any MAPT item. The first time it encountered the item was in the prompting. Therefore, the performance depended solely on the LLM's ability in mathematical reasoning and understanding of the item calibration task. Given this task's demanding nature, we used one of the most capable LLMs, GPT-4 (Hurst et al., 2024).

An example of a prompt is.

Your task is to assess the difficulty value of a given multiple-choice math question, ranging from [min b] to [max b]. Assign a higher difficulty value to more challenging questions. Before outputting the difficulty value, provide a detailed analysis based on the following factors:

1. Mathematical Concepts Involved: Describe the specific mathematical principles and theories that are required to understand and solve this question.

2. Computational Complexity and Student Familiarity: Assess the difficulty level of the calculations required to solve the question, and consider the typical student's familiarity with these types of calculations and computational methods at their current educational level.

3. Clarity and Potential for Misleading or Tricky Wording: Evaluate the clarity of the question's wording and answer choices, and discuss any elements that might confuse test-takers or lead them toward incorrect interpretations or answers.

Use these criteria to inform your evaluation before assigning the final difficulty value. The last line should be: The difficulty value of the target question is: xxx, where xxx is a number.

Question: Which expression represents 4 less than twice a number, n?

Option A: 4 – n.

Option B: n – 4.

Option C: 4 - 2n.

Option D: 2n – 4.

Method 3: LLM fine-tuning.

The third method involved fine-tuning an LLM to predict the difficulty value using the math item text as input. We used RoBERTa (Liu, 2019) as the pre-trained LLM for this method. RoBERTa is an expansion of Google's original BERT (bidirectional encoder representations from transformers) architecture to include all of the web's unannotated text in the training of the LLM. We called this method "fine-tuning a LLM" because we allowed RoBERTa to learn over a training set of math items and their corresponding difficulty values and MAPT levels. The LLM converted the input text to a vector that contained various feature information of the input text, for example, semantic and syntactic features. On a high level, the model learns to represent the meaning of text as a sequence of latent scalar values within a vector output.

A linear regression layer was then applied to the feature vector to predict the difficulty value: this is the primary function that is being adapted in this current architecture. An example of an input is:

Given the math multiple-choice question and its corresponding test level, your task is to evaluate the difficulty score of the math multiple-choice question. The difficulty score ranges from ⌈min b⌉ to ⌈max b⌉, where higher values indicate greater difficulty.

Question: Which expression represents 4 less than twice a number, n ?

Option A: 4 – n.

Option B: n – 4.

Option C: 4 - 2n.

Option D: 2n – 4.

The test level of this math multiple-choice question is 5.0.

2.3. Experimental Setup and Evaluation Metrics

Since our dataset size was small, we used 5-fold cross-validation. This sampling procedure partitioned the dataset into five equal groups and selected a group (that had not been used before) as the testing set, while the other groups served as the training set. We trained and evaluated linear regression and LLM fine-tuning under this configuration. Since CoT prompting did not involve any model training, we only performed evaluation over the folds. This procedure was repeated a total of 5 times with 5 independent evaluation metrics. Finally, we averaged these results. By implementing 5-fold cross-validation, we also reduced the probability of variability in the results related to poor train-test splits.

The evaluation metrics included R-squared, Mean Squared Error (MSE), and the percentage of pairs that maintain the same order as the ground-truth pairs (MATCH). For the third metric, we compared the predicted

difficulty of each pair of test questions to determine whether the first question was more difficult than the second one. We then calculated the percentage of pairs where the predicted order matches the ground-truth orders.

3. RESULTS

We first report results for linear regression, which allowed us to evaluate the value added by the different types of predictor variables in the prediction. After that summary, we present the results across the three approaches.

3.1. Linear Regression Method

Table 1 presents the proportion of variance accounted for (R2) in the IRT difficulty values by specific sets of predictor variables. Of course, the performance increased as we included more predictors. That said, it is interesting to note the most dramatic improvement occurred upon inclusion of the MAPT Level as a feature. R2 increased 45% above that predicted by only the linguistic and math features, which indicates the item developers did a good job increasing the complexity of the items as grade level increased. Additionally, we observed a reduction in MSE of about 0.55 and an increase in MATCH accuracy of about 16% in the same context, which indicates MAPT Level is a good feature for improving prediction accuracy. Adding the additional content information about the items (i.e., Domain Group, Cognitive Level) increased R2 by about 6%.

Table 1. Proportion of variance accounted for in linear regression by specific predictor sets.

Predictors (Features)	R^2	MSE	Match
Linguistics, Math	0.115	1.002	0.639
Linguistic, Math, MAPT level	0.568	0.455	0.796
Linguistic, Math, MAPT level, domain group, cognitive level	0.629	0.417	0.810

3.2. Comparison Across Methods

A comparison of the evaluation metrics across the three methods is presented in Table 2. The linear regression and fine-tuning methods had similar results and vastly outperformed CoT prompting. Fine-tuning an LLM nudged past the R2 of about 63% observed for linear regression by a trivial amount (0.06%). The MSE and MATCH were similar for linear regression and fine-tuning. This result illustrates that fine-tuning an LLM was capable of extracting meaningful features that had a similar effect in predicting the difficulty parameter as those manually defined by educational experts from raw text input. Moreover, the results from fine-tuning an LLM suggested that it understood the scale of the difficulty parameter via training.

Table 2. Comparison of evaluation metrics across methods.

Evaluation metrics	Linear regression	Finetuning	Chain-of-thought
R-squared	0.629	0.635	-0.642
MSE	0.417	0.414	1.85
Match	0.81	0.806	0.684

CoT prompting with GPT-4o had the lowest performance by a large margin. A negative R-squared score indicates that the model performed worse than a constant function that always predicts the mean value of the output, in our case, IRT b-parameters. Our intuition behind this result is that GPT-4 struggled to capture the context of the IRT b-parameter scale. In other models, the output b-parameter is contextualized and influenced by seen and known b-parameters from the training data. Because GPT-4 is the only model that did not have a

dedicated training portion, its understanding of the scale from which the b-parameter is calibrated is severely limited.

Table 3 presents summary metrics and the corresponding histogram (Figure 1) over the first fold in our sampling procedure, we observed that the range of the actual b-parameters exceeds the ranges predicted by our models. Specifically, finetuning was the closest to the actual range, followed closely by linear regression. Chain-of-thought with GPT-4o predicted a minimum b-parameter closest to the true value; however, its maximum predicted b-parameter was the farthest from the true value. The actual mean of the b-parameters was 0.527 when evaluated over the first fold. Linear regression provided the closest estimate to this mean, with finetuning following closely. Chain-of-thought with GPT-4o had the least accurate mean, primarily due to an over-representation of items with negative difficulty values. The actual median of the b-parameters was 0.694 over the first fold. Linear regression was the most effective model for capturing this metric, with finetuning again close behind. Chain-of-thought with GPT-4o was the least accurate, mainly because of an over-representation of items with negative difficulty values.

Table 3. Summary of actual vs predicted b parameters over all models (First fold).

Metrics	Actual difficulty	Linear regression	Chain-of-thought	Finetuning
		Predicted difficulty	Predicted difficulty	Predicted difficulty
Minimum	-1.614	-0.965	-1.5	-1.385
Maximum	2.485	2.405	1.1	2.307
Range	4.099	3.369	2.6	3.692
Mean	0.527	0.594	-0.441	0.649
Median	0.694	0.727	-0.5	0.850

In Figure 1 we observed that fine-tuning represented the predicted range of b-parameters closest to the actual range. Linear regression best captured the "shape" of the distribution for difficulty values above -1, but struggled to represent items with difficulties closer to -2. CoT with GPT-4o over-represented the distribution of items at the lower difficulty range and predicted poorly for items at the higher difficulty range.

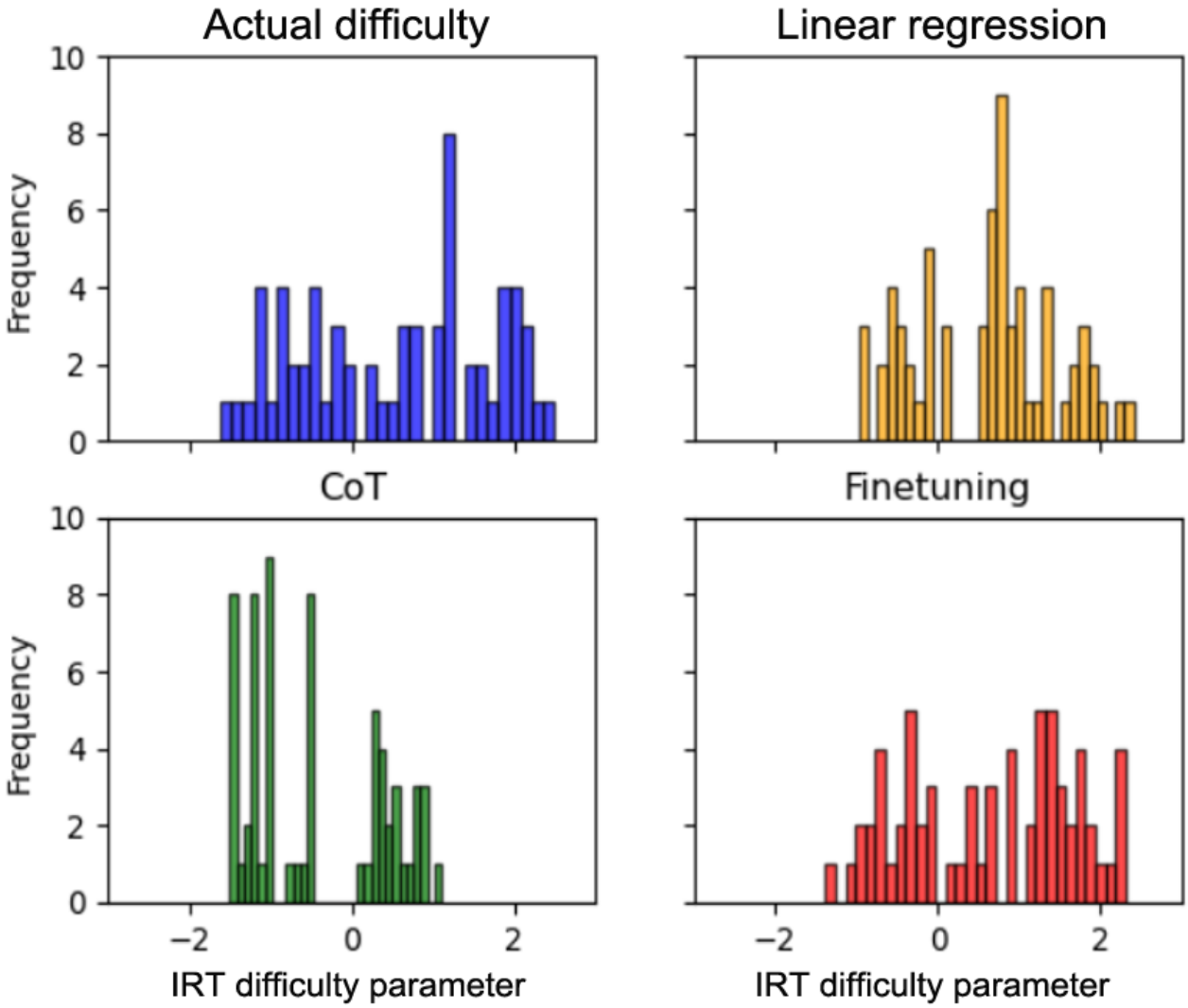

Figure 1. Ground truth difficulty vs predicted difficulty values over all models (First fold).

To summarize our results, the evaluation metrics in Table 2 suggest finetuning and linear regression were almost identical in terms of performance. The summary statistics (Table 3) and Figure 1 showed trade-offs one must consider when choosing the best model. Linear regression was best suited for capturing the "shape" of the distribution for items that had difficulty values above -1. Accordingly, it had the closest maximum, mean, and median to the actual evaluation set. However, linear regression did not capture the distribution of items with difficulty below -1, whereas finetuning performed better. Thus, finetuning was best suited to capture the full range of difficulty values for items. Accordingly, it had the closest minimum and range to the actual evaluation set. However, it is not as accurate as linear regression in capturing the actual distribution for difficulty values closer to 2. Chain-of-thought with GPT-4o was the least competitive model.

4. DISCUSSION

In this study, we proposed and evaluated three methods for accounting for item complexity based on item features. For two of the methods, our results were impressive given the short list of item features we coded, and this was our first foray into the use of AI for this purpose. Although our results fell short of the 81% of variance in task difficulty predicted by Sheehan and Mislevy (1990), our coding scheme represented that of two university researchers and three graduate students (as compared to the resources of a large testing company). We believe that

more time in ascertaining item features via expert analysis would improve the regression results. Likewise, we believe that providing additional input features for fine-tuning will enhance the AI results. One interesting finding is the similarity of the AI-based fine-tuning results with the (manual) linear regression results. As fine-tuning improves, the need to manually code features is likely to diminish. However, such qualitative interpretation of item features is still likely to be necessary to interpret and describe the item complexity scale.

Although this research is in its infancy, the educational implications are significant. If we can model item difficulty using content rather than students' responses, we will no longer depend on pilot-testing items on students for item calibration purposes. Calibrating items based on test-takers' responses is not only time-consuming and costly, but in many cases, the data are contaminated by students who are unmotivated, guessing, or exhibiting various levels of homogeneity, and may not even represent the population to be tested eventually (not to mention disruptions due to catastrophes like COVID-19). It would also ground the interpretation of test scores within the construct domain measured, rather than within a norm-referenced context; thus, improving the validity of test score interpretations.

4.1. Limitations and Future Study

Like all studies, ours had limitations. One limitation is that we excluded items with images as input data to our models. Therefore, the generalizability of these results to math items containing figures and graphs is limited. Future research should incorporate multimodal models to capture the meaning of images within text. Additionally, future research should explore the application of these methods for quantifying the complexity of items in other subject areas, such as reading and science.

Another limitation of our model with GPT-4o was that it did not have a dedicated training portion. GPT-4o struggled with understanding the scale of the difficulty parameter. For future work, we could improve the performance of GPT-4o by adding in-context examples as demonstrations of the task (Brown et al., 2020). This could potentially help GPT-4o better "understand" the scale of the difficulty parameter.

Although LLM fine-tuning achieved similar performance to linear regression, we could improve the training step by "instruction-tuning" (Zhang et al., 2023) RoBERTa with various kinds of math-related tasks. This step could further improve RoBERTa's mathematical understanding and reasoning ability, helping the LLM to estimate the difficulty value of the math items more accurately. Additionally, we might experiment with more powerful open-sourced LLMs, for example, Llama3 (Dubey et al., 2024) to improve performance.

An important finding of this study is that feature selection had a crucial impact on the performance of linear regression. Clearly, further research is needed on features that reflect item complexity. Our goal is to extract construct-relevant features from items to predict difficulty. This goal is in part achieved through many of our features, such as the number of words, the number of propositions, etc. Future work could inform us of new features that better represent the qualities of an item and move us closer towards scaling for 'complexity'.

5. CONCLUSION

AI has the power to realize the theories purported by Fischer (1973) and Whitely (1983) and others to represent the complexity characteristics of items on the test score scale. The present research indicates that LLM fine-tuning may be the most promising approach, but more research is needed to identify construct-relevant attributes to feed the model so it can better reflect the complexity of the tasks the items present to students. Although current IRT item parameters are not a perfect criterion for

evaluating these new scaling approaches, there should be a strong correspondence between the two with respect to the distribution of item difficulty/complexity.

REFERENCES

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, *34*(3), 862-914. https://doi.org/10.1007/s40593-023-00362-1

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*. Paper presented at the 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. (2020). *R2DE: A NLP approach to estimating IRT parameters of newly generated questions*. Paper presented at the Proceedings of the 10th International Conference on Learning Analytics & Knowledge (pp. 412–421).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., & Hu, G. (2019). *DIRT: Deep learning enhanced item response theory for cognitive diagnosis*. Paper presented at the Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2397–2400). ACM.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., & Ganapathy, R. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359-374. https://doi.org/10.1016/0001-6918(73)90003-6

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., . . . Radford, A. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Ling, T., Kang, B. H., Johns, D. P., Walls, J., & Bindoff, I. (2008). *Expert-driven knowledge discovery*. Paper presented at the Proceedings of the Fifth International Conference on Information Technology: New Generations (ITNG 2008) (pp. 174–178). IEEE.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, *364*.

Lord, F. M. (1952). *A theory of test scores*. Psychometric Monographs, Whole No.7.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, *27*(3), 255-272.

Sireci, S. G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, *40*(1), 7-16. https://doi.org/10.1111/emip.12415

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27-51.

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation*, *1*(1), 5. https://doi.org/10.59863/GPML7603

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*(7), 433.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179. https://doi.org/10.1037//0033-2909.93.1.179

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, *26*(3), 163-174. https://doi.org/10.1080/10627197.2021.1956897

Yaneva, V., Baldwin, P., & Mee, J. (2019). *Predicting the difficulty of multiple-choice questions in a high-stakes medical exam.* Paper presented at the Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 11–20). Association for Computational Linguistics.

Yaneva, V., Baldwin, P., & Mee, J. (2020). *Predicting item survival for multiple-choice questions in a high-stakes medical exam.* Paper presented at the Proceedings of the 12th Language Resources and Evaluation Conference (pp. 6812–6818). European Language Resources Association.

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O'Donnell, F., Wells, C. S., . . . Garcia, A. (2024). *Massachusetts adult proficiency tests for college and career readiness technical manual, volume 2.* Retrieved from Center for Educational Assessment Research Report No. 1002. Amherst, MA: Center for Educational Assessment:

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., & Wang, G. (2023). Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792.*