Enhancing Item Difficulty Prediction in Large-scale Assessment with Large Language Model

Mubarak Mojoyinola[1], **Olasunkanmi James Kehinde**[2], **Judy Tang**[3]
[1]Psychological and Quantitative Foundations, University of Iowa, Iowa City
[2] Department of Psychology, Norfolk State University
[3]Survey Management and Design, Westat
[1]mubarak-mojoyinola@uiowa.edu, [2]ojkehinde@nsu.edu, [3]JudyTang@westat.com

Abstract

Item difficulty prediction remains a critical challenge in large-scale assessment development, particularly for international programs like TIMSS where extensive pretesting is costly and time-consuming. This study investigated the utility of large language model (LLM)-extracted cognitive features for predicting item difficulty in mathematics assessment. We analyzed restricted-use TIMSS mathematics items from Grades 4 and 8, comparing three XGBoost models: traditional features (metadata and textual complexity), LLM-extracted cognitive features, and a combined approach. Traditional features alone achieved moderate performance (R^2 = 0.36), while LLM-extracted cognitive demand variables showed weaker individual performance (R^2 = 0.20). However, the combined model substantially outperformed both individual approaches, explaining 48% of variance in item difficulty, a 33% improvement over traditional methods alone. Results demonstrate that LLM-extracted features provide complementary predictive information that enhances difficulty prediction when integrated with conventional textual and metadata features. This approach offers a scalable alternative to expert-based cognitive analysis while maintaining theoretical grounding in established assessment frameworks.

1 Introduction

The calibration of item difficulty is a foundational and resource-intensive requirement in any assessment development. For large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS), the cost is even higher as items need to be pretested across multiple education systems participating in the program (von Davier et al., 2024). The conventional psychometric process relies on extensive field testing to gather empirical data, a practice that presents significant logistical and financial burdens, thereby creating a bottleneck in the test development lifecycle. This operational challenge has motivated a sustained search for automated methods capable of predicting item difficulty directly from item text, aiming to augment and streamline the item development process (AlKhuzaey et al., 2024). Early research in this area leveraged traditional Natural Language Processing (NLP) techniques to extract surface-level and engineered linguistic features. These included readability indices (e.g., Flesch-Kincaid), word and sentence counts, syntactic complexity metrics, and psycholinguistic features from tools like Coh-Metrix (AlKhuzaey et al., 2024). While valuable, these methods do not capture the deeper conceptual, cognitive, and reasoning demands embedded within an assessment item.

The advent of Large Language Models (LLMs) represents a paradigm shift in this domain. Three dominant strategies for using LLMs for item difficulty prediction have emerged. The first, direct estimation, involves prompting an LLM to act as a subject-matter expert and assign a holistic difficulty rating to an item. While intuitively appealing, this approach functions as a "black box," providing a score without rationale, and has shown inconsistent performance, particularly for items designed for younger learners (Razavi and Powers, 2025). The second approach involves using LLMs to generate text embeddings from item stems and response options, which then serve as features in a machine learning model. This approach has been shown to produce accurate difficulty predictions (Bulut et al., 2024; Kapoor et al., 2025). However, models based on embeddings lack interpretability. The third and most sophisticated strategy treats the LLM as a feature extractor. In this two-stage process, the LLM is guided by a structured prompt to analyze an item and output values for a set of predefined, interpretable cognitive and linguistic features. These features are then used as predictors in a separate, often simpler, machine learning model (Razavi and

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 218–222

Powers, 2025).

This study builds upon the promising feature-based methodology by addressing key limitations in current LLM applications for item difficulty prediction. While previous research has demonstrated the potential of LLM-extracted features, significant gaps remain in applying these approaches to complex, international assessment contexts such as TIMSS mathematics items that span multiple grade levels and cognitive domains. Furthermore, existing studies have primarily focused on either traditional psychometric features or LLM-derived metrics in isolation, without systematically investigating how these complementary approaches can be integrated to enhance predictive accuracy. This investigation addresses these limitations by developing a comprehensive framework that combines traditional textual and metadata features with LLM-extracted cognitive demand variables, providing empirical evidence for the added value of automated cognitive feature extraction in large-scale assessment contexts.

Research Questions

1. How do large language model-extracted cognitive features compare to traditional textual complexity features in predicting item difficulty in large-scale assessments?

2. To what extent do LLM-extracted cognitive demand features enhance item difficulty prediction when combined with traditional features?

2 Related Works

Research on item difficulty modeling has long emphasized the integration of psychometric and cognitive frameworks. Sheehan and Mislevy (1994) applied tree-based regression analyses to link item features, solution processes, and response formats with IRT parameters, explaining up to 36% of the variance in difficulty. Competency-based approaches have been particularly influential in international large-scale assessments. Turner et al. (2013) demonstrated that six mathematical competencies (e.g., reasoning, modeling, symbol use) strongly predicted item difficulty in PISA. Similarly, Schneider et al. (2013) showed that Depth of Knowledge (DOK), reading load, and contextual demands systematically predicted item difficulty. With the rise of machine learning, feature-based approaches have advanced prediction. Štěpánek et al. (2023) compared multiple algorithms and found that elastic net and random forests outperformed expert ratings, suggesting that textual features can approximate empirical difficulties. Yi et al. (2024) extended this work with an XGBoost-SHAP framework, achieving strong predictive accuracy while offering interpretability by quantifying the contribution of features such as reasoning steps and symbolic complexity. While these studies demonstrated the importance of cognitive and textual features, they relied on experts to manually code the item data for cognitive features, thereby limiting the number of items that could be studied efficiently.

3 Method

3.1 Dataset

The dataset for this investigation comprised 202 restricted-use mathematics test items selected from TIMSS Grades 4 and 8 assessments administered in 2015 and 2019. These items spanned nine mathematical content areas, with the Number domain contributing the highest number of items, while roughly 23% were eTIMSS items delivered through digital platforms. Item difficulty was quantified using international average proportion-correct values (p-values) obtained from the TIMSS International Database. The mean proportion correct was 51.63%.

3.2 LLM

Using OpenAI's GPT-4.1, several cognitive features were extracted from each item. Leveraging the reasoning capacity of GPT-4.1, we instructed the model through few-shot prompting to evaluate the items and provide appropriate rating based on the provided detailed rubrics with numerical scales and specific criteria for consistent rating across items. GPT-4.1 was accessed through OpenAI's APIs using the ellmer package in R

3.3 Item Features

This study examined variables that could be systematically categorized based on their extraction methodology: traditional features derived from conventional computational and metadata approaches, and LLM-extracted features leveraging large language model capabilities for automated item analysis.

Traditional Features encompass two subcategories of variables that have been extensively used in prior research on item difficulty modeling.

Metadata Variables refer to characteristics specified during item development and assessment administration. These included grade (4 or 8), item type (multiple choice or constructed response), content domain (number, algebra, geometry, data and probability), cognitive domain (knowing, applying, reasoning), and presence of visual elements. These variables align with established TIMSS framework specifications and capture basic structural features of assessment items.

Textual Complexity Variables captured the linguistic demands of item stems and response options using established computational linguistics approaches. These included basic text statistics such as character count, word count, sentence count, and syllable count, as well as established readability indices including the Automated Readability Index, SMOG readability formula, Coleman-Liau index, Flesch Reading Ease, and Gunning Fog index. Additional variables measured the frequency of digits in item stem and response options, recognizing that mathematical text presents unique processing demands beyond general readability.

LLM-Extracted Features represent a novel approach to automated item difficulty modeling, leveraging the reasoning capabilities of large language models to extract features that traditionally required expert human judgment.

Mathematical Content Features were extracted by prompting the LLM to identify and categorize abstract mathematical concepts present in each item. The LLM was also asked to provide a difficulty rating on a 0-100 scale based on its analysis of the mathematical content and cognitive demands, serving as an AI-generated difficulty estimate.

Cognitive Demand Variables were extracted using the LLM to rate items according to the four cognitive competencies framework developed by Turner et al. (2013) for PISA assessment. The LLM was prompted to evaluate each item's demand for: reasoning and argumentation, problem solving, mathematical modeling, and communication. These competencies describe essential cognitive processes required for successful mathematical problem solving and have demonstrated high predictive validity in prior research, with competency-based variables explaining approximately 70% of the variance in PISA item difficulty when used in regression models (Turner et al., 2013). The LLM-based extraction approach offers a scalable alternative to expert panel ratings while maintaining theoretical grounding in established cognitive frameworks.

3.4 Modeling

Using the proportion correct as our estimate of item difficulty, we built a tree-based ensemble model to map item features to difficulty estimates. Specifically, we employed Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), a machine learning algorithm that has demonstrated superior performance across diverse prediction tasks, including item difficulty prediction (Yi et al., 2024; Lamgarraj et al., 2024).

XGBoost operates through a sequential ensemble approach, iteratively constructing weak decision trees where each subsequent model focuses on correcting the prediction errors made by the previously constructed models. This additive modeling strategy allows the algorithm to capture complex, non-linear relationships between item features and difficulty that traditional linear regression approaches might miss. The sequential nature of the boosting process enables the model to progressively refine its predictions by learning from residual errors, ultimately producing a highly accurate composite predictor.

The model's ability to provide feature importance rankings offers valuable insights for assessment development. By quantifying the relative contribution of different item characteristics to difficulty prediction, XGBoost can inform item writers about which features most strongly influence item difficulty, potentially improving the efficiency of item development processes. This interpretability is particularly valuable when comparing the predictive utility of traditional features versus novel LLM-extracted features.

The dataset was randomly partitioned into training (80%) and testing (20%) sets to enable robust model evaluation. Hyperparameter optimization for the XGBoost model was conducted using 5-fold cross-validation with grid search on the training set, ensuring that model selection decisions were based on generalizable performance rather than overfitting to specific data partitions.

Model performance was evaluated using two metrics: root mean squared error (RMSE) to quantify the magnitude of prediction errors, and coefficient of determination (R^2), to quantify the proportion of variance in item difficulty explained by the

model.

4 Results

Three XGBoost models were developed to evaluate the predictive utility of different feature categories for item difficulty prediction: a text-based model using traditional features, a cognitive model using LLM-extracted features, and a comprehensive model combining both feature types. Table 1 presents the performance metrics for all three models based on the testing dataset.

Table 1: Model performance metrics

Feature	RMSE	R^2
Traditional	15.43	0.36
LLM Cognitive	17.27	0.20
Traditional + LLM Cognitive	14.03	0.48

The text-based model, utilizing traditional metadata and textual complexity variables, demonstrated moderate predictive performance on the test set (R^2 = .36, RMSE = 15.43). This model effectively captured the linguistic and structural characteristics of assessment items that influence difficulty, including readability indices, word counts, and domain specifications.

In contrast, the cognitive model using only LLM-extracted features showed weaker individual performance (R^2 = 0.20, RMSE = 17.27). While the cognitive demand variables and mathematical content features extracted by the LLM captured theoretically important aspects of item difficulty, these features alone were insufficient for accurate difficulty prediction.

The combined model incorporating both traditional and LLM-extracted features substantially outperformed either individual approach, achieving the highest test set performance (R^2 = 0.48, RMSE = 14.03). This represents a 33% improvement in explained variance compared to the text-only model and a 140% improvement compared to the cognitive-only model. The superior performance of the integrated approach demonstrates that LLM-extracted cognitive features provide unique predictive information that complements traditional item characteristics.

Figure 1 displays the relative importance of the top ten features for predicting mathematics item difficulty according to the best-performing XGBoost model that combined traditional and LLM-extracted features. The LLM rating of item difficulty emerged as the most important predictor, followed by item type and a series of readability indices.

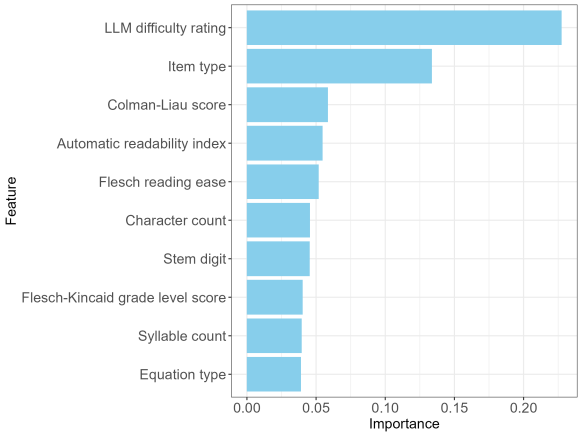

Figure 1: Feature Importance Plot

5 Discussion

The findings from this study demonstrate that integrating LLM-extracted features with traditional item characteristics yields clear benefits for predicting difficulty in TIMSS mathematics assessments. Traditional features alone achieved moderate accuracy (R^2 = 0.36), a result consistent with earlier research that relied on readability indices and metadata such as item format and domain (e.g., Sheehan and Mislevy (1994); Schneider et al. (2013)). LLM-extracted features on their own, while grounded in cognitive frameworks similar to those emphasized by Turner et al. (2013), showed weaker predictive performance, suggesting that automated cognitive coding alone is not sufficient to achieve excellent predictive performance. However, the model with combined feature types explained 48% the variance in item difficulty, surpassing both approaches in isolation and aligning with the findings of Štěpánek et al. (2023) and Yi et al. (2024), who showed that hybrid models incorporating multiple feature sets outperform single source predictors. These results suggest that LLM-derived cognitive measures capture unique dimensions of difficulty that complement rather than replace traditional text and metadata indicators.

The study also reinforces and extends findings from recent LLM-based work. Razavi and Powers (2025) demonstrated that LLMs can enhance difficulty prediction when used as feature extrac-

tors for tree-based models, while Li et al. (2025) found that fine-tuned smaller models often outperform large general-purpose LLMs in educational contexts. Our results resonate with these studies by showing that raw LLM predictions are insufficient but that their structured cognitive features add significant value when combined with traditional descriptors.

6 Conclusion

This study provides new evidence on the potential of LLM-extracted features to improve item difficulty prediction in international large-scale assessments such as TIMSS. While traditional metadata and textual complexity variables accounted for a moderate proportion of variance in item difficulty, and LLM-extracted cognitive features alone showed limited predictive value, their integration substantially enhanced accuracy, explaining nearly half of the variance in item difficulty. These results confirm that LLMs capture complementary aspects of cognitive demand and reasoning that extend beyond conventional text-based measures, offering a scalable alternative to manual coding. The findings also reinforce prior evidence that hybrid models outperform single-source predictors and demonstrate that combining psychometric, linguistic, and cognitive perspectives is essential for advancing item modeling.

By carefully choosing theoretically grounded cognitive demand features, this research shows how LLM-extracted features can provide not only stronger predictions but also actionable insights into the cognitive and structural elements driving item difficulty. Together, these contributions respond to calls in prior work for approaches that balance predictive power with interpretability, bridging psychometric traditions with modern NLP advances. Ultimately, the study offers a practical pathway for improving efficiency in item calibration, reducing reliance on costly pretesting, and enhancing the design of equitable and cognitively grounded assessments in mathematics education.

References

S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item difficulty and response time prediction with large language models: An empirical analysis of usmle items.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Radhika Kapoor, Sang T Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W Domingue. 2025. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*.

Mohamed Lamgarraj, Céline Joiron, Aymeric Parant, and Gilles Dequen. 2024. Exploring item difficulty prediction: Data driven approach for item difficulty estimation. In *International Conference on Intelligent Tutoring Systems*, pages 415–424. Springer.

Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, page 00131644251344973.

P. Razavi and S. J. Powers. 2025. Estimating item difficulty using large language models and tree-based machine learning algorithms. *arXiv preprint arXiv:2504.08804*.

M Christina Schneider, Kristen L Huff, Karla L Egan, Margie L Gaines, and Steve Ferrara. 2013. Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, 18(2):99–121.

Kathleen Sheehan and Robert J Mislevy. 1994. A tree-based analysis of items from an assessment of basic mathematics skills.

Lubomír Štěpánek, Jana Dlouhá, and Patrícia Martinková. 2023. Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19):4104.

Ross Turner, John Dossey, Werner Blum, and Mogens Niss. 2013. Using mathematical competencies to predict item difficulty in pisa: A meg study. In *Research on PISA: Research outcomes of the PISA Research Conference 2009*, pages 23–37. Springer.

M. von Davier, B. Fishbein, and A. Kennedy, editors. 2024. *TIMSS 2023 Technical Report (Methods and Procedures)*. Boston College, TIMSS & PIRLS International Study Center, Boston.

Xifan Yi, Jianing Sun, and Xiaopeng Wu. 2024. Novel feature-based difficulty prediction method for mathematics items using xgboost-based shap model. *Mathematics*, 12(10):1455.