Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Automated Cognitive Feature Generation: LLM Applications in Item Difficulty Modeling

Mubarak O. Mojoyinola *

Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City, Iowa, USA
*Corresponding author. Email: mubarak-mojoyinola@uiowa.edu

Abstract

This study explores the potential of Large Language Models (LLMs) to automate cognitive feature generation for item difficulty modeling (IDM) in educational assessment. Traditional IDM approaches rely on cognitive features manually coded by trained raters, limiting scalability. While recent studies have utilized Natural Language Processing to extract textual features, cognitive features remain challenging to automate. Using Claude 3.5 Sonnet with zero-shot prompting, we generated six cognitive features for 714 Grade 3-5 mathematics items: number of unknown parameters, computation steps, relative definition of unknowns, problem type, equation type, and depth of knowledge. Combined with text-based features (e.g., word counts, readability indices, digit counts), we predicted item difficulty using random forest models. Model incorporating both feature types achieved superior performance (R^2 = 0.10) versus text-based (R^2 = 0.04) or cognitive features (R^2 = 0.04) alone. Feature importance analysis revealed word character count, depth of knowledge, total character count, Flesch reading ease, and stem word count as most impactful. These findings demonstrate LLMs' potential to generate theoretically grounded cognitive features traditionally requiring extensive human effort, streamlining IDM while enhancing its theoretical foundation for more efficient and scalable assessment development.

Keywords: Item difficulty modeling, Large language models, Cognitive features, Mathematics assessment

1. Introduction

Item difficulty modeling (IDM) aims to establish relationships between item features and item psychometric properties such as difficulty and discrimination. This endeavor is crucial in the item development process for large-scale assessments, as these relationships can be exploited to predict the parameters of newly developed items, support content validity evidence, and enhance item calibration in small samples (Graf & Fife, 2012). Early IDM studies employed cognitive features based on theoretical models that explain why items vary in difficulty based on required cognitive resources and identify which steps or sub-processes are likely to present challenges to examinees (Graf & Fife, 2012). For mathematics items, examples of cognitive features explored in predicting item difficulty include the number of unknown parameters to be solved, sources of equations needed to solve problems, computation count, and the need to manipulate variables (Daniel & Embretson, 2010; Enright et al., 2002).

These cognitive features require manual coding by raters, which constrains the number of items that can be included in such studies. In order to automate the feature extraction process and enable the application of more complex statistical models for parameter prediction, researchers have leveraged natural language processing (NLP) to extract textual features (features based on item text) (AlKhuzaey et al., 2023). Since NLP techniques cannot be readily utilized for extracting cognitive

features, modern IDM studies mostly rely on linguistic features extracted from item text and/or text representations through embeddings.

However, the importance of different item feature categories depends on the specific goals of an IDM study. When the goal is to maximize parameter prediction accuracy or to obtain auxiliary information for use as priors in item calibration processes, any features that achieve these objectives are valuable, regardless of their interpretability (Belov et al., 2025; Ulitzsch et al., 2025). In contrast, when IDM is used to support validity evidence or inform the item development process, interpretable features, particularly cognitive features, take priority.

Large Language Models (LLMs) have shown significant potential in psychometric research, yielding promising results in automated item generation, item pretesting, and the automated scoring of texts and mathematical graphs (Hao et al., 2024). While LLMs have been incorporated into item difficulty modeling studies, their applications have remained limited in scope. Current approaches primarily employ LLMs to generate text embeddings from item content (Bulut et al., 2024) or to calculate the proportion of items that LLMs answer correctly (Rogoz & Ionescu, 2024). Researchers then combine these LLM-derived metrics with traditional text-based features as predictors in item difficulty models. However, this approach fails to leverage LLMs' capacity for extracting theoretically grounded cognitive features that capture the underlying psychological processes required to solve assessment items.

This study presents a novel methodological approach that employs LLM to automate the extraction of theoretically grounded cognitive features from mathematics assessment items, addressing the traditional bottleneck of manual expert coding that has constrained the scalability of cognitive feature-based item difficulty modeling. Through zero-shot prompting with Claude 3.5 Sonnet, we systematically generate six cognitive features based on established cognitive theory frameworks, then integrate these LLM-derived features with traditional text-based features in a random forest prediction model. This approach offers a scalable solution for incorporating cognitive theory into automated item difficulty prediction while maintaining the interpretability essential for validity evidence and item development processes.

1.1 Research Questions

1. Can LLM generate cognitive features?
2. Which feature type (text features, cognitive features, or a combination of both) most accurately predicts item difficulty?

2. Methods

2.1 Data

The dataset used in this study comprises 714 Grade 3–5 mathematics items from released accountability assessments across five states and NAEP, curated by Beiting-Parrish (2022). The items span seven mathematical domains: Geometry, Measurement and Data, Number and Operations (Fractions), Number and Operations in Base Ten , Number Properties and Operations, Operations and Algebraic Thinking, Statistics and Probability, and The Number System. Each item includes the stem, response options, and p-values as difficulty indices. Table 1 presents a detailed summary of item distributions and p-values.

2.2 Features

We employed text-based features and generated six cognitive features that are based on previous research.

Table 1. Descriptive Statistics of P-values by Grade

		P value		
Grade	N	Min	Mean	Max
3	207	0.27	0.65	0.94
4	285	0.23	0.62	0.95
5	222	0.28	0.61	0.92

2.2.1 Textual Features

Textual features were derived to quantify the linguistic complexity of item stems and response options using established computational linguistics methods. We extracted basic text statistics, including counts of characters, words, sentences, and syllables, along with widely used readability measures such as the Automated Readability Index, SMOG readability formula, Coleman-Liau index, Flesch Reading Ease, and Gunning Fog index. Given that mathematical content imposes distinct cognitive processing requirements beyond conventional readability, we also quantified the prevalence of numerical digits in both item stems and response options. In total, fourteen text-based features were automatically extracted from item stems and response options using the R package `sylcount` (Schmidt, 2023). These features comprised word, sentence, character, and syllable counts for item stems; the aforementioned readability indices; digit counts for stems and response options; and a binary indicator identifying whether the stem contained more digits than the response options.

2.2.2 Cognitive Features

Cognitive features characterize the mental processes and knowledge components required for successful item attempt. Six cognitive features were identified based on prior research (Embretson & Daniel, 2008; Schneider et al., 2013; Yi et al., 2024) and extracted using zero-shot prompting with the Claude 3.5 Sonnet large language model (Anthropic, 2024), accessed through the R package `ellmer` (Wickham et al., 2025).

Number of Unknown Parameters. This feature quantifies how many variables or quantities the examinee must solve for in a given problem. Items requiring the solution of multiple unknowns impose greater cognitive demands because examinees must track multiple pieces of information simultaneously and coordinate their solution strategies across these variables. For instance, a problem asking students to find both the length and width of a rectangle presents greater complexity than one asking only for the perimeter. Previous research has established that the number of unknowns is a significant predictor of item difficulty in mathematics assessment (Embretson & Daniel, 2008).

Number of Computational Steps. This feature captures the number of distinct mathematical operations or procedures required to arrive at the correct answer. Items demanding multiple computational steps require examinees to execute a longer sequence of operations, increasing both the working memory load and the opportunity for errors at each step. A problem requiring students to first convert units, then multiply values, and finally round the result involves more computational steps than one requiring only a single multiplication operation. Research has consistently shown that computational complexity contributes substantially to item difficulty (Daniel & Embretson, 2010; Embretson & Daniel, 2008).

Relational Definition of Unknowns. This binary feature indicates whether unknown parameters in the problem are defined in relation to one another. When unknowns are relationally defined (e.g., Sarah has three times as many apples as John), examinees must mentally construct and maintain the relationship between variables, adding an additional layer of cognitive processing beyond simply identifying what needs to be solved. In contrast, independently defined unknowns can be processed separately without requiring examinees to coordinate relationships between variables.

This relational complexity has been identified as a key source of difficulty in mathematical problem solving (Embretson & Daniel, 2008; Morrison & Embretson, 2014).

Problem Type. This feature distinguishes between abstract mathematical problems and those embedded in real-world contexts. Abstract problems present mathematics in symbolic or purely numerical form (e.g., Solve for x: 2x + 5 = 13), while real-world problems situate mathematical operations within practical scenarios (e.g., "Maria bought 2 notebooks for $5 each. If she paid with a $20 bill, how much change did she receive?"). Real-world problems may impose additional cognitive demands related to comprehension of the context and translation of verbal information into mathematical form (Yi et al., 2024).

Equation Type. This categorical feature identifies the source and form of equations needed for problem solution, with four possible values: (1) no equation needed, (2) equation explicitly given in the problem, (3) equation must be translated from the problem statement, or (4) equation must be recalled from memory. Items requiring examinees to translate verbal descriptions into algebraic equations or to recall appropriate formulas from memory impose greater cognitive demands than those providing equations explicitly or requiring no formal equations. This feature captures both the retrieval demands and the translation processes central to mathematical problem solving (Daniel & Embretson, 2010; Embretson & Daniel, 2008).

Depth of Knowledge (DOK). This feature categorizes items according to Webb's Depth of Knowledge framework, which characterizes the cognitive complexity required for successful performance across four levels: Level 1 (recall and reproduction), Level 2 (skills and concepts), Level 3 (strategic thinking), and Level 4 (extended thinking). Higher DOK levels require increasingly sophisticated cognitive processes, from simple retrieval of facts or procedures to complex reasoning, planning, and extended chains of inference. The DOK framework has been widely adopted in educational assessment to describe the cognitive demands of test items and has demonstrated strong relationships with empirical item difficulty (Schneider et al., 2013).

A sample prompt for extracting equation type is provided below:

```
system prompt = "You are an expert in assessment development and
educational measurement analyzing mathematics items for {grade}
students."

user prompt = "Which of these options applies to the equation needed
to solve this question: (1) No Equation Needed, (2) Given Equation,
(3) Translate Equation, and (4) Recall. Return
only the number associated with each option. {Question}"
```

2.3 Machine Learning Model

Item difficulty was predicted using a random forest model with different combinations of feature sets. The random forest algorithm builds multiple decision trees from bootstrap resamples of the training data and combines their predictions through averaging, which minimizes variance and mitigates overfitting. This approach effectively captures nonlinear patterns and feature interactions, and has demonstrated strong performance in item difficulty prediction studies (Štěpánek et al., 2023; Xu et al., 2022). We employed an 80/20 data partitioning strategy, allocating 80% of the dataset for model training and reserving 20% for testing. During training, 5-fold cross-validation combined with grid search was utilized for hyperparameter tuning. Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

3. Results

Table 2 presents the model's performance on the test data for different combinations of item features. The model incorporating both text-based and cognitive features achieved the best predictive performance, while models using only text-based or cognitive features showed similar predictive performance. Thus, combining both types of features offers an improvement over using only one type. Although the resulting R^2 is low, which is typical in most IDM studies (Ferrara et al., 2022), the combination of text-based and cognitive features resulted in a 150% increase in R^2 compared to using only one type of feature. This finding aligns with results obtained by (Belov et al., 2022), as presented in (Ulitzsch et al., 2025), where using a combination of expert-created features and text features led to a substantial increase in R^2.

Table 2. Performance Metrics on Test Data

Feature	RMSE	MAE	R^2
Text	0.15	0.12	0.04
Cognitive	0.15	0.12	0.04
Text + Cognitive	0.14	0.12	0.10

3.1 Feature Importance

Feature importance based on mean decrease in node impurity was examined for the best-performing random forest model, which combined text-based and LLM-extracted cognitive features. This metric quantifies each feature's contribution to the model's predictive performance. Figure 1 shows that word character count, LLM-extracted depth of knowledge, total character count, and Flesch reading ease are the strongest predictors of item difficulty.

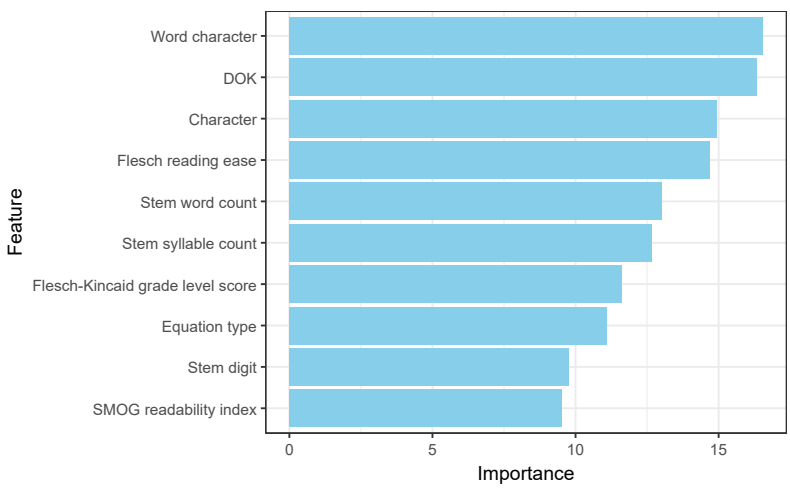

Figure 1. Variable importance rankings for the top ten features.

4. Discussion

This study demonstrates that Large Language Models can effectively automate the extraction of cognitive features from mathematics assessment items, addressing a longstanding scalability challenge in item difficulty modeling research. Using Claude 3.5 Sonnet with zero-shot prompting, we

successfully extracted six theoretically grounded cognitive features across 714 items without requiring training examples or fine-tuning. This finding extends previous LLM applications in item difficulty research, which have primarily focused on text embedding (Bulut et al., 2024; Kapoor et al., 2025), by demonstrating their utility in feature extraction tasks that traditionally required expert human judgment.

Our results indicate that combining text-based and cognitive features yields superior predictive performance compared to using either feature type alone. The model incorporating both feature types achieved an R^2 of 0.10, representing a 150% improvement over models using only text-based or cognitive features (R^2 = 0.04 for each). This finding aligns with prior research demonstrating that combining expert-created features with text-based features substantially enhances predictive accuracy (Belov et al., 2022; Ulitzsch et al., 2025). The additive value of cognitive features suggests that they capture aspects of item difficulty not fully reflected in surface-level linguistic characteristics.

Feature importance analysis revealed that word character count, depth of knowledge, total character count, and Flesch reading ease were the strongest predictors. Notably, depth of knowledge emerged as the second most important feature overall and the most influential cognitive feature. This finding underscores the value of incorporating theory-based cognitive features, as DOK levels explicitly operationalize cognitive complexity frameworks that motivated early item difficulty modeling research (Graf & Fife, 2012).

Several limitations warrant consideration. The modest R^2 values, while consistent with typical findings in item difficulty modeling research (Ferrara et al., 2022), indicate that substantial variance remains unexplained. This likely reflects the multifaceted nature of item difficulty, which encompasses response processes and item characteristics not captured in our feature set. Additionally, we did not directly compare LLM-generated features against expert human ratings, and our study employed only zero-shot prompting with a single LLM. Future research should validate LLM-generated features against expert ratings, explore different prompting strategies and LLMs, and examine generalizability to other content areas.

Despite these limitations, our findings have important practical implications. By automating cognitive feature extraction, LLMs can substantially reduce the time and resources required for item difficulty modeling while preserving the theoretical grounding that makes such models interpretable and useful for supporting validity evidence. This capability is particularly valuable for large-scale assessment programs that develop hundreds or thousands of items annually, where manual cognitive coding would be prohibitively expensive. This study provides initial evidence that LLMs offer a scalable solution for incorporating cognitive theory into automated item difficulty prediction, with potential applications across various assessment contexts and content domains.

Acknowledgement

I sincerely acknowledge Dr. Maggie Beiting-Parrish for her valuable work in curating the dataset used in this study across multiple sources and making it publicly available.

Competing Interests None

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.

Anthropic. (2024). Claude 3.5 sonnet [Large language model developed by Anthropic]. https://www.anthropic.com/claude

Beiting-Parrish, M. (2022). *Analyzing the linguistic features of standardized math items: A text mining approach* [Doctoral dissertation, City University of New York].

Belov, D. I., Topczewski, A., & McVay, A. (2022). Predicting irt 3pl parameters via a neural network model. *International Meeting of the Psychometric Society*.

Belov, D. I., Lüdtke, O., & Ulitzsch, E. (2025). A supervised learning approach to estimating irt models in small samples. *British Journal of Mathematical and Statistical Psychology*.

Bulut, O., Gorgun, G., & Tan, B. (2024). Item difficulty and response time prediction with large language models: An empirical analysis of usmle items.

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, *34*(5), 348–364.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science*, *50*(3), 328.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied measurement in education*, *15*(1), 49–74.

Ferrara, S., Steedle, J. T., & Frantz, R. S. (2022). Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education*, *35*(3), 237–253.

Graf, E. A., & Fife, J. H. (2012). Difficulty modeling and automatic generation of quantitative items: Recent advances and possible next steps. *Automatic item generation*, 157–178.

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative ai. *Educational Measurement: Issues and Practice*, *43*(2), 16–29.

Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., & Domingue, B. W. (2025). Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*.

Morrison, K. M., & Embretson, S. E. (2014). Using cognitive complexity to measure the psychometric properties of mathematics assessment items. *Multivariate Behavioral Research*, *49*(3), 292–293.

Rogoz, A.-C., & Ionescu, R. T. (2024). Unibucllm: Harnessing llms for automated prediction of item difficulty and response time for multiple-choice questions. *arXiv preprint arXiv:2404.13343*.

Schmidt, D. (2023). *Sylcount: Syllable counting and readability measurements* [R package version 0.2-6]. https://CRAN.R-project.org/package=sylcount

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

Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, *11*(19), 4104.

Ulitzsch, E., Belov, D., Lüdtke, O., & Robitzsch, A. (2025). Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test. *Journal of Educational Measurement*.

Wickham, H., Cheng, J., & Jacobs, A. (2025). *Ellmer: Chat with large language models* [R package version 0.1.1]. https://CRAN.R-project.org/package=ellmer

Xu, J., Wei, T., & Lv, P. (2022). Sql-dp: A novel difficulty prediction framework for sql programming problems. *International Educational Data Mining Society*.

Yi, X., Sun, J., & Wu, X. (2024). Novel feature-based difficulty prediction method for mathematics items using xgboost-based shap model. *Mathematics*, *12*(10), 1455.