

Predicting IRT Parameters for Passage-Based Reading Items Using NLP

Ruitao Liu

ACT

ruitao.liu@act.org

Abstract

This study predicts IRT item difficulty and discrimination for passage-based reading comprehension items using lexical, syntactic, and semantic NLP features. Modeling interactions among passages, stems, and options, Light-GBM with semantic embeddings best predicts difficulty ($r = 0.594$), while discrimination remains harder to recover from text alone.

1 Introduction

Item Response Theory (IRT) supports item calibration, test construction, equating, and score interpretation in large-scale educational assessment (Embretson and Reise, 2000; Lord, 1980). In operational practice, however, item parameters are not available until an item has been administered to a sufficiently large sample. During early stages of item development, item writers, reviewers, and psychometricians therefore make important decisions without direct evidence about how a new item is likely to function. This creates strong interest in methods that can approximate likely item behavior before field data become available.

Passage-based reading comprehension items are a particularly important case. Such items depend on more than a short question and several answer options. Examinees must read an extended passage, understand the task defined by the stem, and distinguish the correct answer from distractors that may be partially plausible. Difficulty may arise from vocabulary burden, syntactic complexity, discourse integration, or fine-grained semantic contrasts among options. Discrimination may depend even more strongly on how well distractors compete with the correct answer for students at different ability levels. These conditions make passage-based reading items both practically valuable and methodologically demanding for text-based prediction.

Recent work has shown that textual features can be informative for difficulty prediction. Stud-

ies in reading and educational NLP have used text-complexity measures, machine learning, and embedding-based representations to model item difficulty from content (Sheehan et al., 2010; AlKhuyaey et al., 2024; Štěpánek et al., 2023). The area has accelerated further with the BEA shared task on automated prediction of difficulty and response time for multiple-choice questions and with LLM-based pipelines for educational assessments (Yaneva et al., 2024; Veeramani et al., 2024). Yet much of this work focuses on short items or domains where extended contextual reading is not central.

The present study addresses that gap by focusing on long-passage reading items and by explicitly modeling interactions among passage, stem, and answer options. The paper makes a practical as well as methodological contribution. Practically, it is motivated by real operational work and is intended to support future item development workflows. Methodologically, it compares lexical, syntactic, and semantic feature families under form-level cross-validation and examines variable importance for the strongest text-only model. The main finding is that item difficulty is moderately recoverable from text alone, especially when semantic interaction is represented well, whereas item discrimination remains substantially harder to predict.

2 Related Work

The present study is related to two research traditions. The first is work on explanatory item features and text complexity. In reading assessment, automated measures such as SourceRater showed that lexical frequency, syntactic burden, and discourse-related signals can be aligned with human judgments of text complexity (Sheehan et al., 2010). This work did not directly target calibrated IRT parameters for full passage-based multiple-choice items, but it established the broader premise that

linguistic properties of text are relevant to assessment difficulty.

The second tradition is direct prediction of question or item difficulty from text. AlKhuyaey et al. (2024) reviewed automatic text-based approaches and showed that machine-learning methods now dominate this area. More directly relevant to the present paper, Štěpánek et al. (2023) compared multiple algorithms for predicting difficulty from English reading-comprehension item wording and showed that textual features can compete with domain-expert judgments. The BEA 2024 shared task further demonstrated broad interest in automated difficulty prediction, while Veeramani et al. (2024) illustrated how newer LLM-based pipelines can be combined with linguistic features for improved estimation.

What remains less studied is the case of passage-based reading assessment, where the item is not a single short prompt but a coordinated textual system. Work on scenario-based reading assessment emphasizes that modern reading tasks require integration across extended context, task instructions, and answer choices (Sabatini et al., 2014, 2020). This perspective matters for item-parameter prediction because a model that only inspects surface wording in the stem or isolated option text may miss the deeper interaction structure that gives the item its difficulty. The present study therefore treats passage, stem, and options as distinct but interdependent sources of signal.

3 Data and Study Design

The analysis begins with an operational data pipeline that parses assessment XML files, extracts passage and item text, and merges those textual fields with psychometric records containing calibrated item parameters. For each item, the resulting dataset includes an extended passage, a stem, four answer options, a correct key, and operational estimates for item difficulty and item discrimination. Items are grouped by form, and the present analysis uses eight reading forms after cleaning and alignment checks. The current phase is deliberately restricted to text-derived inputs so that the contribution of linguistic information can be evaluated in isolation from metadata or behavior logs.

Two targets are modeled separately. The first is item difficulty, represented by the operational b parameter. The second is item discrimination, represented by the operational a parameter. Separate

models are trained for the two targets because they reflect different psychometric functions. Difficulty is more closely related to the overall processing burden of the item, whereas discrimination depends on how strongly the item differentiates among examinees across the ability scale.

The evaluation design follows the operational use case. Each form is held out in turn as a test fold, while the remaining forms are used for training. This form-level cross-validation prevents leakage across items from the same form and gives a more realistic picture of how the models would behave on a truly new form. Performance is reported with root mean squared error (RMSE), R^2 , and Pearson correlation, which together capture absolute error, explained variance, and rank-order agreement.

4 Feature Engineering and Modeling

4.1 Lexical Features

Lexical features describe surface properties of language in the passage, stem, and answer options. The feature set includes token counts, mean and median ZIPF word frequency based on SUBTLEX norms (van Heuven et al., 2014), part-of-speech-specific frequency summaries, and normalized ratios that distinguish absolute from relative lexical difficulty. The features are computed separately for the major text components because the source of lexical burden may differ across them. A passage may be difficult because it contains rare academic vocabulary, while an option may be difficult because it uses unusually precise wording that competes strongly with the correct answer.

To make answer-option information more informative, raw option features are transformed into contrastive variables. For each lexical measure, the value for the correct answer is retained and the three distractor values are ranked and stored separately. This design preserves the full distribution of distractor properties rather than reducing them to a single mean. The same logic is important later for discrimination modeling, because what matters is often not the absolute wording of one option, but the contrast between the correct answer and the most competitive distractor.

4.2 Syntactic Features

Syntactic features capture structural complexity. They include sentence-count and sentence-length summaries, dependency-tree depth, average branching factor, independent and dependent clause

counts, subordinate-clause ratios, phrase-density measures, dependency-relation frequencies, part-of-speech trigram diversity, and heuristic indicators of ambiguity. These features are motivated by well-established reading theory: readers must hold more information in working memory when sentences are longer, more deeply embedded, or structurally ambiguous. The same principle can apply to answer options, where syntactic structure influences perceived naturalness and plausibility.

As with lexical features, syntactic information is represented both at the component level and at the contrastive choice level. This makes it possible to capture cases in which the correct answer is structurally simpler than all distractors, or in which one distractor is made especially persuasive by a complete clause structure. Because passage-based reading items often combine long contextual text with short options, separating passage-level and option-level structure is especially useful.

4.3 Semantic Interaction Features

Semantic features form the core of the present study. We use transformer-based sentence embeddings, principally all-mpnet-base-v2 and all-MiniLM-L6-v2, to place passage sentences, stems, and options in a shared semantic space (Reimers and Gurevych, 2019). From those representations we derive several families of features. Passage-sentence embeddings are pooled with mean, median, minimum, maximum, and standard-deviation summaries. Stem-option cosine similarities measure how directly an option aligns with the task stated in the stem. Correct-versus-distractor margins measure how much the correct option separates from the strongest competitor. Option-option clustering features summarize how tightly distractors group together. Passage-option distances and projections along the stem-passage axis model whether an option is better understood as a passage-supported inference, a stem-level paraphrase, or a semantically misleading distractor.

These semantic features are important because long-passage items are not solved through local matching alone. A student may need to connect information from several sentences, interpret the discourse goal of the question, and reject distractors that are partially supported by isolated fragments of the text. Lexical and syntactic features can reflect processing burden, but semantic interaction features are more directly tied to the construct of reading comprehension. They therefore provide a

principled way to represent passage–stem–option interaction, which is the main conceptual emphasis of this paper.

4.4 Models and Evaluation

We compare a constant baseline, Elastic Net regression (Zou and Hastie, 2005), and LightGBM (Ke et al., 2017). The baseline predicts the mean parameter value from the training forms and serves as a minimum reference. Elastic Net provides a linear model with regularization and feature selection, which is useful in high-dimensional spaces. LightGBM provides a non-linear model that can exploit thresholds and feature interactions. We evaluate lexical-only, syntactic-only, semantic-only, syntactic-plus-semantic, and full lexical-plus-syntactic-plus-semantic configurations in order to understand which kinds of information matter most.

5 Results

5.1 Item Difficulty

Table 1 summarizes the main form-level results. For item difficulty, the baseline yields mean RMSE = 0.964. A semantic-only LightGBM model already reduces RMSE to 0.794 and reaches mean $R^2 = 0.300$ and mean correlation = 0.581. Adding syntactic features produces the best overall text-only result, with mean RMSE = 0.788, mean $R^2 = 0.313$, and mean correlation = 0.594. Adding lexical features on top of that produces no further gain, which suggests that once strong semantic information is available, lexical signal contributes relatively little incremental value for this target.

The difficulty results also show meaningful variability across forms. For the best text-only model, fold-wise RMSE ranges from 0.621 to 1.007, while fold-wise correlation ranges from 0.410 to 0.816. The median correlation is 0.552, with an interquartile range from 0.488 to 0.707. Thus, although predictive performance differs across forms, the consistently positive correlations indicate that the model captures signal that transfers to held-out forms. At the same time, the observed variation suggests that some forms are more difficult to predict than others, motivating further investigation of form-level differences and sources of prediction error. The stronger performance of LightGBM relative to Elastic Net further indicates that non-linear interactions matter for difficulty prediction.

Model	RMSE	R^2	Corr.
Difficulty baseline	0.964	0.000	—
Difficulty	0.794	0.300	0.581
LightGBM, semantic only			
Difficulty	0.788	0.313	0.594
LightGBM, syntactic + semantic			
Discrimination	0.266	0.000	—
Discrimination	0.247	0.066	0.501
Elastic Net, semantic only			
Discrimination	0.247	0.082	0.500
Elastic Net, full features			

Table 1: Selected form-level cross-validation results for the main baseline and best text-only models.

5.2 Item Discrimination

Item discrimination is markedly harder to predict. The baseline yields mean RMSE = 0.266. The best text-only models are all Elastic Net variants based on all-mpnet-base-v2 semantic information. The semantic-only model yields mean RMSE = 0.247, mean R^2 = 0.066, and mean correlation = 0.501. The full lexical-plus-syntactic-plus-semantic model is nearly identical in RMSE and correlation, with a small improvement in mean R^2 to 0.082. In contrast to difficulty, LightGBM does not help here; the best non-linear discrimination models remain near baseline in RMSE and produce unstable R^2 values.

The instability of discrimination prediction is visible at the fold level. For the best semantic-only Elastic Net model, fold-wise R^2 ranges from -0.838 to 0.355, and fold-wise correlation ranges from 0.193 to 0.677. The median correlation is 0.586, but some forms are clearly much harder than others. This pattern is substantively plausible. Discrimination depends not only on text quality but also on how different groups of students respond to the distractors, and those interaction effects are only partly observable from text alone.

5.3 Variable Importance

To understand why the best difficulty model works, we analyzed gain-based variable importance from the LightGBM model trained on the syntactic-plus-semantic feature set with all-mpnet-base-v2 em-

Feature group	Share of gain
Option embeddings	44.1%
Pooled passage sentence embeddings	41.5%
Stem embeddings	13.4%
Syntactic features	0.9%
Explicit semantic interaction summaries	0.1%

Table 2: Gain-based importance by feature group for the best difficulty model.

beddings. The model contains 8,169 candidate predictors. Of these, 7,727 are semantic features and 438 are syntactic features. Gain is concentrated overwhelmingly in the semantic part of the representation. Option embedding dimensions account for 44.1% of total gain, pooled passage sentence embeddings account for 41.5%, and stem embeddings account for 13.4%. All syntactic features together account for only 0.9%, while explicitly coded semantic interaction summaries contribute roughly 0.1%.

The top individual predictors are also revealing. The single highest-gain feature is a pooled passage-sentence mean embedding dimension, followed by a pooled passage-sentence median dimension and two option-embedding dimensions. This pattern indicates that the model relies most on global passage meaning and the semantic positioning of answer options, with stem meaning playing a secondary role. The low gain assigned to hand-engineered interaction summaries does not mean that interaction is unimportant. Rather, it suggests that the tree ensemble can recover interaction structure directly from raw passage and option embeddings when those representations are sufficiently rich.

6 Discussion and Operational Implications

Several conclusions follow from these results. First, difficulty is more recoverable from text than discrimination. This makes psychometric sense. Difficulty is closely tied to overall linguistic burden and integration demand, which are visible in passages, stems, and options. Discrimination, by contrast, depends on the fine structure of distractor competition within a population of examinees. Text contributes to that process, but it does not fully determine it. Second, semantic signal is far more important than lexical signal for passage-based reading items. Lex-

ical features alone are too shallow for long passages in which the main challenge may be inference, integration, or subtle semantic contrast rather than rare words alone.

These findings have direct operational value. A text-based model of difficulty is not a replacement for field testing, but it can still support early stages of item development. It can help flag items that are likely to be unusually hard, identify passages whose semantic structure may create excessive burden, and provide one additional input to pretesting and form assembly decisions. The present results are particularly encouraging because they are obtained without response-process variables, timing information, or hand-coded metadata. In other words, the models already recover meaningful signal from text alone.

At the same time, the paper is appropriately framed as work in progress. The current system is intentionally text-only, and several important steps remain before operational use. We are extending the feature space with newer embedding models and LLM-based representations, investigating whether discourse-aware and reasoning-aware features improve performance, and testing hybrid models that combine text with metadata or response-based variables. We are also working on interpretability, because operational users need more than accurate predictions; they need explanations that can inform item revision. Finally, validation beyond the current reading data is necessary before stronger claims about generality can be made.

7 Conclusion

This paper examined whether IRT item parameters for passage-based reading comprehension items can be predicted from text alone. Using items from eight operational reading forms, we found that text-only models provide meaningful prediction for item difficulty, with the strongest model reaching mean RMSE = 0.788, mean R^2 = 0.313, and mean correlation = 0.594 under form-level cross-validation. Item discrimination proved substantially harder to model, although semantic Elastic Net models still improved modestly over baseline. Variable-importance analysis showed that the most useful signal is concentrated in passage and option embeddings, which supports the view that long-passage item difficulty is driven mainly by semantic integration across item components. The results suggest a realistic role for AI-assisted item

analysis in operational assessment while also defining clear boundaries for what text-only prediction can currently achieve.

References

- Samah AlKhuyaey, Floriana Grasso, Terry R. Payne, and Valentina A. M. Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Susan E. Embretson and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3146–3154.
- Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.
- John P. Sabatini, Tenaha O'Reilly, Laura Halderman, and Kelly Bruce. 2014. Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges. *L'Année Psychologique*, 114(4):693–723.
- John P. Sabatini, Tenaha O'Reilly, Jonathan P. Weeks, and Zuowei Wang. 2020. Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, 20(1):1–23.
- Kathleen M. Sheehan, Irene Kostin, Yoko Futagi, and Michael Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, 2010(2):i–44.
- 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.
- Walter J. B. van Heuven, Pia Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Hariram Veeramani, Surendrabikram Thapa, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Large language model-based pipeline for item difficulty and response time estimation for educational assessments. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 561–566.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 470–482.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.