Predicting Item Difficulty and Generating Reading Comprehension Items via an Annotated Repository

Radhika Kapoor, Mayank Sharma, Sang T. Truong, Maria Araceli Ruiz-Primo, Nick Joseph Haber, Benjamin W. Domingue
Stanford University, USA
{rkap786,masharma,sttruong,aruiz,nhaber,bdomingu}@stanford.edu

Abstract

Prediction of item difficulty from its text content is of substantial interest for automated generation of test items. In this paper, we focus on the related problem of recovering IRT-based difficulty when the data originally reported item p-value (percent correct responses). We model this item difficulty using a repository of reading passages and student data from US standardized tests from New York and Texas for grades 3-8 spanning the years 2018-23. This repository is annotated with meta-data on (1) linguistic features of the reading items, (2) test features of the passage, and (3) context features. Using a penalized regression model, we achieve an RMSE of 0.59 (compared to a 0.92 baseline) and a 0.77 correlation between true and predicted difficulty. We further evaluated the impact of LLM embeddings (ModernBERT, BERT, and LLaMA), finding that they marginally improve performance but function effectively as standalone alternatives to traditional linguistic features. Finally, we demonstrate how this difficulty prediction model powers a publicly available, human-in-the-loop tool for generating reading comprehension items.

1 Introduction

Education systems use reading comprehension tests to evaluate student achievement and growth, to monitor average school and teacher performance, and to diagnose students for learning disorders. Creating items for these tests can be an expensive and time-intensive endeavor (Case and Swanson, 1998; Rudner, 2009), especially for standardized tests in K-12 schools. Conventionally, subject matter experts (e.g., teachers or reading specialists) create the items used in standardized tests (Attali et al., 2014). Items are then piloted in schools and student response data is collected to estimate item properties like difficulty and to confirm that items are measuring student knowledge as intended (Lane et al., 2016, Chapter 1).[1] However, items developed by subject matter experts might not perform as expected when taken to the field (Bejar, 2012; Reid, 1991; Impara and Plake, 1998; Dee and Domingue, 2021).

Education researchers have long been interested in modeling the difficulty of test items. Improved item difficulty models could support subject matter experts in creating items as well as reduce field testing time and costs (Bejar, 1991; Gorin and Embretson, 2006). However, this field has achieved mixed results, as modeling and predicting item difficulty is complex, and requires advanced text analysis and annotation to extract relevant item features (Gorin and Embretson, 2006).

Large Language Models (LLMs) have recently shown the potential to generate items with desirable characteristics like difficulty and to automate scoring of text-based responses (Ferrara et al., 2018; Huang et al., 2017; White et al., 2022; Attali et al., 2022; Zelikman et al., 2023) possibly reducing the time and cost associated with these tasks (Acharya et al., 2023). In this paper, we leverage LLMs with item and respondent features to predict mean difficulty of reading comprehension items. We build a repository of items using reading comprehension items from New York and Texas, administered to Grades 3-8 students spanning 5 years between 2018 and 2023. This repository contains the item text (reading passage, question, and distractors), and the state-level average percent correct responses (p-value). Items are annotated with (1) linguistic features of the reading passage, including their lexical, semantic, and syntactic characteristics and readability metrics (2) test features of the item, such as the presence of highlighted or bold text (3) respon-

[1] New York: https://www.nysed.gov/state-assessment/nysed-test-development-process; Texas: https://tea.texas.gov/texas-schools/accountability/academic-accountability/performance-reporting/8-why-does-texas-field-test.pdf

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 777–797

dent's state, grade, and year (4) LLM-generated text embeddings for reading comprehension items.

We focus on how effectively LLM-generated embeddings predict item difficulty in an attempt to understand whether transformer-based models can augment traditional psychometric methods used to generate items. Our models achieve an RMSE of 0.59 compared to a baseline of 0.92, and a correlation of 0.77 between true and predicted item difficulty. Previous results for predicting difficulty using item content achieved correlation of 0.3-0.7, and smaller RMSE improvements (Ha et al., 2019; Huang et al., 2017). We make the following contributions to literature: First, to our knowledge, this is the first publicly available work to examine automatic difficulty modeling for reading items for standardized tests in US schools. Second, we also convert mean p-value of items from different grades to a vertical logit scale using a Rasch IRT model, which could be used in future work to aggregate response data collected from different populations. Third, we compare difficulty prediction from different types of item features: linguistic features, test features, pre-trained LLM embeddings, and respondent context. We report similar results for prediction from linguistic features or LLM embeddings, suggesting that any of these classes of metrics could be used by themselves for prediction. Fourth, we compare the performance of three LLMs (BERT-base, ModernBERT, and LLaMA), as well as different prompts for generating LLM embeddings. We report similar results for these variations. Finally, we introduce a publicly available tool designed for practitioners and researchers to generate reading comprehension items with predicted difficulty. The tool can be found at https://item-generator-production.up.railway.app/.

2 Background

2.1 Features of Reading Comprehension Items

A reading comprehension item in a test is usually structured as a passage that the respondents read, followed by a question or prompt, and then by several options from which respondents select one response. The item might also include a few other components: directions about the passage or the item (e.g., "read the passage and answer the questions that follow"), images or figures (e.g., to highlight part of the story), and tables (e.g., with data related to the story). This section describes the features of reading comprehension items that could predict their difficulty.

2.1.1 Features of Reading Passages

Reading standards such as the Common Core Standards for English Language Arts & Literacy[2] are tied to these complexity measures, both for teaching and tests. Text complexity of reading passages has been captured using (1) descriptive metrics, such as the count of words in passage, number of sentences, and number of paragraphs (2) linguistic complexity, which is influenced by factors such as complexity of words (measures include word length and word frequency), syntactic complexity (measures include modifier propositional density or count of adjectives per 1000 words), and text cohesion (defined as words and ideas overlapping throughout the text) (Graesser et al., 2014; McNamara et al., 2014).

Corpus analysis tools can parse text for their linguistic features. A popular corpus analysis tool is the Coh-Metrix, which offers over 100 measures of vocabulary, syntax, semantics, and other linguistic features. Coh-Metrix also reports aggregate measures of text cohesion and readability using metrics like Flesch-Kincaid, latent semantic analysis and principal components analysis (McNamara et al., 2014). Measures from corpus analysis tools, such as length of passage, concreteness or abstractness of text, and vocabulary level have been reported to be correlated with difficulty for reading questions (AlKhuzaey et al., 2023; Choi and Moon, 2020).

In addition to the reading complexity of the passage, its layout or structure might also affect student performance. The layout here refers to reading cues that point students to important aspects of the content (such as text segmentation), the inclusion of boxes or footnotes, directions that point readers to important vocabulary, helpful images, and options for text-to-speech in computer-based reading (Lawrence et al., 2022; Mize et al., 2020).

2.1.2 Features of Items

Item features affect their difficulty by changing their interpretability. Some features can make it easier to locate and retrieve information from text (Lawrence et al., 2022; Le Hebel et al., 2017; Ruiz-Primo and Li, 2015). For example, an item can include relevant sentences from the passage in a

[2] https://achievethecore.org/page/2725/text-complexity

text box or definitions of difficult words. Other features emphasize key aspects of items. For instance, the word "agree" in the item is emphasized: "Which of the following statements would <*name of character*> in the story **agree** with?". In multiple choice questions, the distractors or response options also affect difficulty; for instance, some distractor options could be very similar to the correct answer. The type of content (e.g., use of images) might also influence student performance.

2.2 Item Difficulty Modeling

2.2.1 Automatic generation of items

Automated item generation was first proposed in the 1970s (Gierl and Haladyna, 2013), though the field saw a resurgence in the 2000s with the need for more item creation with computerized tests, as well as with an increase in computational power. A type of automatic item generation uses expert-created templates, item schemes or structures, parts of which can be populated using algorithms (Embretson and Yang, 2006; Kurdi et al., 2020). Such cognitive item model templates are specific to the item type being created.

With recent advances, LLMs can be prompted to generate items without additional tuning or information ("zero-shot" generation) or with some examples in context ("one-shot" or "few-shot" generation) (Attali et al., 2022; Kurdi et al., 2020). LLMs can also be prompted using cognitive models, templates, or rules to create items with desired properties (Gierl and Haladyna, 2013; Kurdi et al., 2020). These templates can specify the type of item to be generated (e.g., fill-in-the-blank questions, multiple-choice questions, open-ended questions), provide further instructions based on item type (e.g., the number of distractors and correct responses for multiple choice items), specify reasoning to be elicited (e.g., recall or inference), and even state the desired item difficulty (e.g., "create an easy item that 70% of respondents would get right") (Sayin and Gierl, 2024). While LLMs can be used to generate big quantities of test items (Kurdi et al., 2020), their quality, especially properties like difficulty, is unknown, making them hard to use reliably in educational testing.

2.2.2 Item difficulty modeling

The problem of item generation is fundamentally linked with the problem of item difficulty modeling. Items are created with an implicit or explicit difficulty level, defined by human experts using heuristics or by statistical modeling. Popular models used for statistical modeling include supervised learning approaches such as linear regression, decision trees, and random forests. Approaches like deep learning and neural nets have also been leveraged but are less common, likely because of the smaller size of education datasets and the desire for inference in addition to model prediction.

Statistical item difficulty modeling approaches usually have the following components: (1) A method to parse item text and images: this can use NLP and image parsers, linguistic corpus analysis tools, and human annotation (2) Model to predict difficulty. Recently, difficulty modeling has been supplemented by embeddings from LLMs (Kurdi et al., 2020; Sharpnack et al., 2024). Embeddings can be extracted from pre-trained models, or models could be fine-tuned. In the shared task organized during the BEA workshop at NAACL 2024 (Yaneva et al., 2024), item difficulty (defined as the percentage of respondents who responded incorrectly) was modeled for 667 MCQ items from the United States Medical Licensing Examination. The best prediction model achieved RMSE of 0.29 compared to a baseline RMSE of 0.31 from a dummy regressor model; the best model used a combination of LLM embeddings and other item features. In a similar exercise with 12,038 MCQ items from the United States Medical Licensing Examination, the best model used a combination of embeddings from Word2Vec and ElMo and linguistic features together, achieving RMSE of 22.45 compared to a baseline RMSE of 23.65.

The results of existing item difficulty prediction efforts have achieved limited success. A key constraint is that education settings rarely provide sufficiently large datasets for fine-tuning LLMs.

3 Data

The dataset contains multiple-choice items from the New York State Testing Program (NYSTP) and Texas STAAR reading comprehension standardized tests for Grades 3-8. NYSTP data is available for the years 2018, 2019, 2022, and 2023, and Texas STAAR data is available for 2019, 2021, and 2022. In total, there are 1076 items based on 170 passages across grades and years from both states (see item counts for grades by state and year in Appendix A).

Features available for passages and related multiple-choice items include item difficulty, defined as mean percent correct response at the state,

year, and grade level. Appendix B reports mean percent correct responses (i.e., the item's *p-value*[3]) across items by grade, state, and year. Note that mean accuracy for both states centers around 60% across grades and years. Student acquisition of additional reading comprehension skills is being offset by increases in overall item difficulty. As a consequence, reported difficulties or p-values cannot be directly compared across grades. In other words, let us say there is a Grade 3 item with p-value 0.60 (i.e., 60% of 3rd graders get this item right) and a Grade 8 item with p-value 0.60. The Grade 3 item is presumably easier than the Grade 8 item, since Grade 8 students are likely to be better readers; however, the p-values reported in the dataset would consider these items to be equally difficult. We convert these p-values to a common vertical logit scale to make them comparable for difficulty modeling. This is described in more detail in the Methods section.

4 Methods

This section describes how the dataset is processed for item difficulty prediction. Section 4.1 describes how mean p-values across grades and years are rescaled to a uniform logit vertical scale. Section 4.2 describes annotation of items with linguistic, test, and context features used as inputs in the prediction model. Section 4.3 describes how embeddings are generated from pre-trained LLMs. Finally, Section 4.4 then describes how these inputs are used to predict item difficulty using a penalized regression model.

4.1 Rescaling p-values

We use publicly available estimates of average student abilities across grades to convert average p-value to a common vertical scale. Appendix C shows the NWEA MAP 2020 reading scale[4], which reports mean grade-level achievement norms for over 500,000 students attending public schools across 50 states. These average scores are used to transform the p-values from a specific grade to a scale that can then be used for direct comparison; for example, on our derived scale the 3rd grade item with a p-value of 0.6 will have a lower value than the 8th grade item with a p-value of 0.6 given the growth from 200.74 to 220.93 observed on the NWEA scale.

[3]Discussed in Chapter 14 in Crocker and Algina (1986)

[4]https://www.nwea.org/uploads/2020/02/NY-MAP-Growth-Linking-Study-Report-2020-07-22.pdf

4.1.1 Rescale p-values to IRT difficulty

To rescale the p-values to the logit scale, we take advantage of a basic feature of 1PL IRT models (von Davier, 2016) that suggests a relationship between such p-values and item-person characteristics:

$$\Pr(y = 1) = p_{ij} = \sigma(\theta_i + b_j) \tag{1}$$

where θ is the relevant person ability, b is the item easiness, and σ is a sigmoid function (the logistic sigmoid, $\sigma(x) = \frac{1}{1+\exp(-x)}$, which we also use here). Note that we predict item easiness rather than item difficulty, to remain consistent with the original dataset, which reports p-values as measures of easiness. We don't have access to the relevant θ_i and b values that generated the data here, but we utilize the fact that they, in particular average θ values by grade level, are publicly reported for similar groups. We thus do the following. We denote the mean ability for students in a given grade (using values in Appendix C) as $\overline{\theta_g}$. For each item j given to students in grade g, we can calculate easiness b_j based on the relationship between the p-value p_j and the above Equation 1. Specifically, we solve for b_j such that

$$p_{ij} = \frac{1}{1 + \exp(-(\overline{\theta_g} + b_j))} \tag{2}$$

$$\iff b_j = -\overline{\theta_g} + \log\left(\frac{p_j}{1 - p_j}\right) \tag{3}$$

We use the resulting b_j values in our subsequent modeling.

4.1.2 Adjustment Results

In the analysis below, $\overline{\theta_g}$ is drawn from the NWEA Spring 2020 reading vertical scale. Figure 1 shows the results of the adjustment process. The top panel shows grade-level mean b_j values for NY and Texas, mapped from p-values using $\overline{\theta_g}$ from NWEA 2020 reading achievement scores. The bottom panel shows the distribution of the unadjusted and adjusted p-values by grade. The unadjusted mean p-values (bottom left panel) are all distributed around 0.6 with no increasing pattern as grade levels increase. This pattern is adjusted on the logit scale in the bottom right panel. Note that adjusted Grade 3 items are easiest on average, while Grade 8 items are hardest. Hence, an average p-value of 0.6 for Grade 3 corresponds to 0.3 on the adjusted scale, and maps to -1.69 for Grade 8.

Note one critical fact. The values on the y-axis in Figure 1 are based on the NWEA Scale in Appendix C. While the NWEA scale is a high-quality

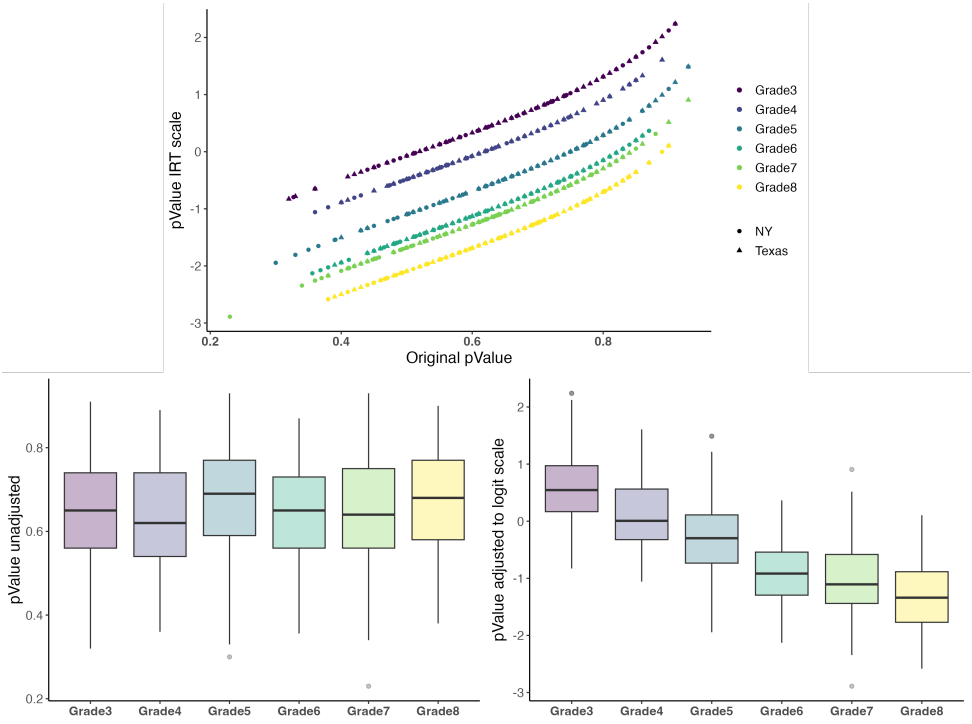

Figure 1: Conversion of p-value to IRT Easiness by Grades

Note. (Top): Conversion of p-value to IRT scale by Grade; (Bottom Left): Unadjusted p-values by Grades; (Bottom Right): Adjusted Mean Easiness by Grades on the Logit Scale

scale derived from large samples, it is one of several scales that could provide a common metric across grades (Dadey and Briggs, 2012). Given that this choice of $\overline{\theta_g}$ is somewhat arbitrary, we consider a range of potentially plausible $\overline{\theta_g}$ values in ancillary robustness analyses (discussed in Appendix G).

4.2 Annotation of item bank

Items are annotated with three categories of features: context, test characteristics, and linguistic text analysis measures.

4.2.1 Context Features

Respondent context characteristics are expected to be some of the most important predictors of difficulty. In this dataset, the context variables are the state, year, and grade of respondents, for which average difficulty values are reported.

4.2.2 Test Features

Test features refer to the presentation or design elements in a test which could influence response. For the passage, this includes the use of highlighted (bolded, italicized, or underlined) text to emphasize important elements. For questions, this refers to the use of highlighted text (e.g., the use of bold text in "6. Which sentence **best** states the main idea of 'Around the World'?"), or inclusion of relevant text or sentences from the reading passage. For example, in the following question, the sentence in bold is repeated from the passage: "Read this sentence from paragraph 6. **'Traveling east across the Atlantic, Nellie took just one bag to move quickly.'** How is this detail important to paragraph 1?". The inclusion of these test characteristics is captured as indicator variables (=1 when given test feature is present in the item). Appendix D lists the variables that capture relevant test characteristics about an item.

4.2.3 Text Analysis Features

Measures of linguistic complexity of reading passages are generated using Coh-Metrix software package and included as predictors for the analysis (Appendix E). The indicators are divided into the following categories: (1) Descriptive metrics, such as count of paragraphs, count of sentences, mean number of words in a sentence, and number of letters or syllables in a word (2) Readability metrics, such as Flesch-Kincaid score and Coh-Metrix L2 Readability (3) Ease of reading or text "easability", estimated through principal components analysis

(PCA) components, measuring syntactic simplicity, vocabulary concreteness, referential cohesion, and deep cohesion (4) Text cohesion, where more cohesive texts are easier to read; measured by word overlap across text or semantic overlap in text (5) Deep cohesion in text, measured by use of causal verbs like *although* and *overall*, (6) Syntactic complexity, measured by use of more complex syntax such as density of noun phrases or syntax similarity of complex sentences (7) Word complexity, based on word frequency or psychological traits of words such as age of acquisition. These measures are generated by the Coh-Metrix software (McNamara et al., 2014), except for the Flesch-Kincaid score (generated by readability package in Python[5]), and count of paragraphs in passage and word count of passage, which was generated in R.

4.3 Sentence Embedding from Large Language Models

LLM-generated embeddings for question text and distractors are used as predictors in difficulty modeling. Distractors are tagged with `correct answer` or `wrong answer` as appropriate. The question text, correct answer and distractors, and passage are merged into one input string, for which embeddings are generated (Appendix F). We experiment with BERT (Devlin et al., 2018), ModernBERT base model with 149M parameters (Warner et al., 2024) and LLaMA-3.1-8B (Dubey et al., 2024) for the generation of sentence embeddings.

To generate embeddings from ModernBERT[6] and BERT[7], each item is converted to a statement by merging the question text and available options as follows: (1) question text (2) a tag for `correct answer`, followed by the correct option (3) a tag for `wrong answer`, followed by the first wrong answer in the question. This is repeated for the remaining wrong answers (4) Reading passages. The objective is to capture interactions between the passage, question text, and distractors. The BERT tokenizer is used to convert these sentences into token IDs. The embeddings are derived from the last hidden state of the model. Specifically, for each sentence, we determine the position of the last meaningful token by identifying the index of the final non-padding token. The hidden state of the final token is extracted. Each sentence produces a tensor whose size depends on the number of tokens times the embedding dimension. Because sentence lengths vary, these tensors are averaged across tokens to produce fixed-size embeddings (768 for BERT and ModernBERT, 4096 for LLaMA). This tensor is the final embedding output for the item. The dimensionality of these embeddings is also compressed using PCA to capture 80% of variation. PCA on embeddings also makes it harder to recover the question text and options from the embeddings, which would be a test security concern for high-stakes tests.

We also experimented with other methods of generating embeddings from BERT by removing the reading passage from the text input to tokenizer. We also generate cosine similarity between `correct answer` and `wrong answers` as new predictor variables. The results have some sensitivity to the method of embedding generation, but main conclusions do not change (Appendix G, Table 9).

4.4 Prediction Model

Item-level difficulty (i.e., b_j from Eqn 3) is predicted through a Ridge (L_2) penalized regression that tuned the regularization penalty parameter (λ)[8] via 5-fold cross-validation (repeated and averaged 5 times with a 80% train, 20% test split). All predictor variables are centered at 0 with standard deviation set to 1. The performance of the prediction models is measured by two indicators: RMSE, and Pearson Correlation Coefficient between the true (y_j) and predicted outcome ($\widehat{y_j}$) variable (AlKhuzaey et al., 2023). RMSE is sensitive to the training dataset, and hence is meaningful only for within-dataset model comparisons. The correlation coefficient, however, is better suited for benchmarking this model against results reported in the literature.

5 Results

The predictors include the following categories of variables: (1) context characteristics of state, grade, and year (2) test characteristics (3) linguistic features of reading passage and item (4) LLM-generated embeddings (either the full set of embeddings or PCA components that capture 80% of the variation). The baseline model for comparison sets the predicted difficulty to the mean difficulty for

[5] https://pypi.org/project/readability/

[6] https://huggingface.co/answerdotai/ModernBERT-base

[7] https://huggingface.co/google-bert/bert-base-cased

[8] Elastic net models that tune the ratio between ridge and lasso (α) always converged to the more parsimonious ridge model

Table 1: Results for Predicting Item Difficulty

	RMSE		Correlation	
	Train	**Test**	**Train**	**Test**
Results from human annotated features				
State, Grade, Year	0.60	0.64	0.74	0.72
Test features	0.89	0.92	0.13	0.05
Text analysis features	0.59	0.62	0.75	0.75
All annotated features and context	0.53	0.59	0.81	0.76
Results from LLM embeddings only				
BERT embeddings	0.60	0.66	0.76	0.71
LLaMA embeddings	0.44	0.66	0.89	0.70
ModernBERT embeddings	0.58	0.62	0.77	0.76
Results from LLM embeddings and annotated features				
Annotated features & BERT embeddings	0.60	0.64	0.76	0.73
Annotated features & PCA on BERT embeddings	0.58	0.64	0.76	0.72
Annotated features & LLaMA embeddings	0.47	0.62	0.87	0.75
Annotated features & PCA on LLaMA embeddings	0.59	0.63	0.75	0.73
Annotated features & ModernBERT embeddings	0.57	0.61	0.78	0.76
Annotated features & PCA on ModernBERT embeddings	0.59	0.63	0.75	0.73
Results from LLM embeddings, annotated features, and context				
All features & BERT embeddings	0.59	0.64	0.77	0.74
All features & PCA on BERT embeddings	0.52	0.60	0.81	0.76
All features & LLaMA embeddings	0.47	0.61	0.87	0.76
All features & PCA on LLaMA embeddings	0.52	0.59	0.81	0.76
All features & ModernBERT embeddings	0.57	0.61	0.78	0.77
All features & PCA on ModernBERT embeddings	0.52	0.59	0.81	0.77

Note. Outcome variable is easiness drawn from IRT scale based on NWEA MAP Reading scale Spring 2020

the item; this baseline RMSE was 0.92. This serves as a comparison point for model performance as measured by RMSE, which are dependent on the sample data set.

5.1 Prediction results from annotated features and embeddings

The best performing model uses all available features and PCA on embeddings from ModernBERT or LLaMA (Table 1). The best RMSE on test data is 0.59 and correlation between true and predicted easiness is 0.77.[9] It is not surprising that the best model uses all available features for prediction. We also note that PCA on embeddings performs marginally better than actual embeddings; this is likely because of the small dataset size compared to the number of predictors.

[9]Appendix G reports the results when other vertical scales are used to convert p-values to IRT easiness: for different scales, RMSE varies between 0.53 and 0.61, and correlation varies between 0.75 and 0.85. The RMSE is 3.24 and the correlation is 0.96 when different scales are used for pre- and post-2020 years. This is not strictly comparable with other results which use the same scales for all years

5.2 Prediction results from LLM embeddings only

Results from only LLM embeddings, without using any respondent context characteristics or linguistic analysis of text, are close to the best model (Table 1). The correlation of true and predicted difficulty is 0.76 for ModernBERT with an RMSE of 0.62. When all additional features are added, model performance increases marginally for RMSE to 0.61 and correlation to 0.77. BERT and LLaMA show slightly bigger improvements in model performance when all features are added: RMSE decreases from 0.66 when only BERT embeddings are included to 0.64 when all features are included; the correlation between true and predicted difficulty in the test dataset increases from 0.71 to 0.74. For LLaMA, RMSE decreases from 0.66 to 0.61, and correlation increases from 0.70 to 0.76.

5.3 Prediction results from annotated features

Model performance for annotated features (context, test, and text analysis) is also close to the best model. This is largely because of the text analysis features, which are able to predict item difficulty, suggesting that LLMs and text analysis features may be capturing similar information. Text

analysis features offer the added benefit of being interpretable by humans and require less computation power, which may be an advantage over using LLMs.

Examining the standardized ridge regression coefficients reveals which features are most important for prediction (Appendix F.2, Table 7). Among text analysis features, the strongest predictors are measures of referential cohesion and syntactic complexity. Specifically, noun phrase overlap between adjacent sentences (CRFANP1) and noun phrase density (DRNP) are the two largest predictors, both negatively associated with item easiness — passages with denser and less cohesive noun phrase structure tend to produce harder items. Among word-level features, higher word frequency (WRDFRQa) and greater imageability of content words (WRDIMGc) are both associated with easier items. Notably, traditional readability metrics such as Flesch-Kincaid rank poorly relative to these deeper linguistic features, suggesting that surface-level readability scores commonly used in educational practice may be insufficient for capturing item difficulty.

The overall prediction power of test features by themselves is low, as would be expected given there are only four such variables. However, one test feature appears among the top predictors overall: the presence of highlighted text in the question (*ques_text_highlight_yn*) is positively associated with item easiness, consistent with prior work suggesting that question-level presentation cues help students locate relevant information in the passage.

It is also important to note that these features can predict difficulty without including context characteristics (state, grade, year). These models could hence be used to predict average item difficulty before extensive field pilots, potentially reducing costs.

5.4 Automatic generation of items

The item difficulty model powers the Automatic Item Generator (`https://item-generator-production.up.railway.app/`), a web-based, human-in-the-loop tool that leverages the Claude API for item creation, along with the option for manual edits by users. This system facilitates the creation of assessment materials by (1) Item Generation: Reading items are generated by the Claude generative AI agent, using the existing item bank as context for generation. (2) Difficulty Calibration: Item difficulty scores are calculated on a logit scale using ModernBERT embeddings.

The tool allows for specific parameter adjustments, including the target grade level, passage length, and the number of distractors. It also supports custom context, enabling researchers to integrate their own item generation manuals. Without additional prompting, the tool generates a synthetic dataset that mirrors the characteristics of standardized test items from New York and Texas.

6 Discussion

This paper aims to predict item difficulty using its text content. This approach bridges the gap between manual item creation and modern computational methods, providing insights into the nuanced interplay between text complexity and student response behaviors. The study underscores the transformative potential of transformer-based models, which can generate meaningful representations of items without fine-tuning or additional context information. By rescaling difficulties using IRT and vertical ability scales, the study addressed the challenge of comparing average difficulty across grades, offering a scalable solution when there is no common population or anchor items for traditional test equating and linking approaches. This item difficulty prediction model could be useful for stakeholders like teachers or testing firms, who would be interested in creating items with desirable item properties like difficulty. For instance, an item generated through generative AI could be assigned a difficulty score using this model; this could then be used to assemble appropriate tests.

Our model demonstrated strong performance in estimating item difficulty, achieving a correlation coefficient of 0.77 between predicted and observed difficulty - a result obtained both by ModernBERT embeddings alone and by models combining all available features with LLM embeddings. To the best of our knowledge, this correlation between true and predicted difficulty is one of the best for multiple-choice reading comprehension test items. The RMSE is high, but we expect that it is a function of the constraints of the dataset; the dataset we use is relatively small and the outcome variable is average item difficulty converted to a logit scale. The best model also reduces RMSE from a baseline value of 0.92 to 0.59. Other work that predicts difficulty as a continuous measure reports Pearson's correlation between 0.38 and 0.4 (Huang et al., 2017) and a reduction in RMSE from 23.65

to 22.45 (Ha et al., 2019).

The best results were obtained for the models that combined text analysis and human generated features with LLM embeddings. Surprisingly, in our dataset, using LLM-based embeddings or text analysis features was close to the best performing model. These findings suggest that LLMs could predict item difficulty without additional linguistic analysis or human annotations. Embeddings were generated from three models, BERT, ModernBERT and LLaMA; the results are also similar for the three LLM models. LLaMA embeddings were computationally more expensive to generate but their model performance was better. ModernBERT achieved the best performance, suggesting it might be best suited to such an item difficulty modeling task. BERT's strong performance despite being the simplest model could be a function of the limited size of the dataset, where variations in embedding generation might not lead to improvements in results.

Limitations

While the findings affirm the utility of automated methods, they also highlight several limitations of static embeddings in fully capturing dynamic student interactions. First, the reliance on state-specific datasets from New York and Texas may limit the generalizability of the findings to other educational contexts or student populations. Second, the embeddings generated from LLMs are dependent on pre-trained models that may not be optimally tuned to K-12 educational texts or standardized test items, potentially missing domain-specific nuances. Lastly, the model's predictions are based on aggregated item difficulty, which may oversimplify individual differences in student abilities or test-taking contexts.

Future research should focus on expanding the dataset to include diverse state tests and international benchmarks, such as PISA or data high-stakes tests, to enhance the robustness and applicability of the model. Additionally, fine-tuning LLMs on domain-specific corpora could further improve their ability to capture educational nuances. A comprehensive evaluation of item difficulty prediction should explore its impact on downstream tasks, such as automated item generation and field testing, to validate the practical utility of the proposed model in educational settings. Last, but perhaps most important, this model could be used by teachers or test creators to get early feedback on item properties.

Ethical Considerations

The use of automated tools in educational assessment raises important ethical considerations. Item difficulty prediction models, including the one presented here, should be understood as decision-support tools rather than replacements for human judgment. Decisions about item inclusion, exclusion, or revision in high-stakes assessments should remain with qualified subject matter experts, psychometricians, and assessment professionals who can evaluate item quality in context.

The automatic item generation tool described in this paper is intended to assist practitioners in creating reading comprehension items more efficiently. However, items generated by any AI system require careful human review before use in consequential testing contexts. Automated generation does not guarantee that items are free from bias, culturally appropriate, or aligned with specific curriculum standards. Practitioners using such tools should be aware that generated items may reflect patterns and assumptions present in the training data, which in this case is drawn from two specific US state assessments. There is also a risk that over-reliance on difficulty prediction models could narrow the range of items developed, if practitioners consistently target items predicted to fall within a specific difficulty range.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. LLM based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2023. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, pages 1–53.

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.

Yigal Attali, Luis Saldivia, Carol Jackson, Fred Schuppan, and Wilbur Wanamaker. 2014. Estimating item difficulty with comparative judgments. *ETS Research Report Series*, 2014(2):1–8.

Isaac I Bejar. 1991. A generative approach to psychological and educational measurement. *ETS Research Report Series*, 1991(1):i–54.

Isaac I Bejar. 2012. Item generation: Implications for a validity argument. In *Automatic item generation*, pages 40–55. Routledge.

Susan M Case and David B Swanson. 1998. *Constructing written test questions for the basic and clinical sciences*. National Board of Medical Examiners Philadelphia.

Inn-Chull Choi and Youngsun Moon. 2020. Predicting the difficulty of efl tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, 17(1):18–42.

Linda Crocker and James Algina. 1986. *Introduction to classical and modern test theory*. ERIC.

Nathan Dadey and Derek C Briggs. 2012. A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, 17(14):n14.

Thomas S Dee and Benjamin W Domingue. 2021. Assessing the impact of a test question: Evidence from the "underground railroad" controversy. *Educational Measurement: Issues and Practice*, 40(2):81–88.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Susan Embretson and Xiangdong Yang. 2006. 23 automatic item generation and cognitive psychology. *Handbook of statistics*, 26:747–768.

S Ferrara, J Steedle, and R Frantz. 2018. Item design with test score interpretation in mind. In *Item difficulty modeling: Lessons learned and future directions. Coordinated session at the annual meeting of the National Council on Measurement in Education, New York, NY*.

Mark J Gierl and Thomas M Haladyna. 2013. *Automatic item generation: Theory and practice*. Routledge.

Joanna S Gorin and Susan E Embretson. 2006. Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30(5):394–411.

Arthur C Graesser, Danielle S McNamara, Zhiqang Cai, Mark Conley, Haiying Li, and James Pennebaker. 2014. Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2):210–229.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 11–20.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

James C Impara and Barbara S Plake. 1998. Teachers' ability to estimate item difficulty: A test of the assumptions in the angoff standard setting method. *Journal of Educational Measurement*, 35(1):69–81.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Suzanne Lane, Mark R Raymond, Thomas M Haladyna, and 1 others. 2016. *Handbook of test development*, volume 2. Routledge New York, NY.

Joshua F Lawrence, Rebecca Knoph, Autumn McIlraith, Paulina A Kulesz, and David J Francis. 2022. Reading comprehension and academic vocabulary: Exploring relations of item features and reading proficiency. *Reading Research Quarterly*, 57(2):669–690.

Florence Le Hebel, Pascale Montpied, Andrée Tiberghien, and Valérie Fontanieu. 2017. Sources of difficulty in assessment: Example of pisa science items. *International Journal of Science Education*, 39(4):468–487.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Minnie Mize, Yujeong Park, Megan Schramm-Possinger, and Mari Beth Coleman. 2020. Developing a rubric for evaluating reading applications for learners with reading difficulties. *Intervention in School and Clinic*, 55(3):145–153.

Jerry B Reid. 1991. Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, 10(2):11–14.

Lawrence M Rudner. 2009. Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing*, pages 151–165. Springer.

Maria Araceli Ruiz-Primo and Min Li. 2015. The relationship between item context characteristics and student performance: The case of the 2006 and 2009 pisa science items. *Teachers College Record*, 117(1):1–36.

Ayfer Sayin and Mark Gierl. 2024. Using openai gpt to generate reading comprehension items. *Educational Measurement: Issues and Practice*, 43(1):5–18.

James Sharpnack, Kevin Hao, Phoebe Mulcaire, Klinton Bicknell, Geoff LaFlair, Kevin Yancey, and Alina A von Davier. 2024. Banditcat and autoirt: Machine learning approaches to computerized adaptive testing and item calibration. *arXiv preprint arXiv:2410.21033*.

Matthias von Davier. 2016. Rasch model. In *Handbook of item response theory*, pages 31–48. Chapman and Hall/CRC.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Julia White, Amy Burkhardt, Jason Yeatman, and Noah Goodman. 2022. Automated generation of sentence reading fluency test items. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Victoria Yaneva, Kai North, Peter Baldwin, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, Brian Clauser, and 1 others. 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 (BEA 2024)*, pages 470–482.

Eric Zelikman, Wanjing Anya Ma, Jasmine E Tran, Diyi Yang, Jason D Yeatman, and Nick Haber. 2023. Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency. *arXiv preprint arXiv:2310.06837*.

A Item Counts by State, Grade, and Year

Table 2: Count of Items for Grades by State and Year

State	Grade	Item counts by year					Total
		2018	2019	2021	2022	2023	
NY	Grade 3	12	12		12	17	53
NY	Grade 4	12	12		12	17	53
NY	Grade 5	21	21		21	19	82
NY	Grade 6	21	21		21	19	82
NY	Grade 7	21	21		21	26	89
NY	Grade 8	21	21		21	26	89
Texas	Grade 3		34	28	34		96
Texas	Grade 4		32	32	32		96
Texas	Grade 5		29	34	34		97
Texas	Grade 6		36	36	36		108
Texas	Grade 7		37	37	38		112
Texas	Grade 8		40	39	40		119
Total		108	316	206	322	124	1076

B Mean Percent Correct Response (p-values) by State, Grade, and Year

Table 3: Mean Percent Correct Response (p-values) by State, Grade, and Year

State	Grade	Mean percent correct by year				
		2018	2019	2021	2022	2023
NY	Grade 3	0.60	0.65		0.65	0.60
NY	Grade 4	0.58	0.57		0.64	0.56
NY	Grade 5	0.62	0.64		0.65	0.60
NY	Grade 6	0.67	0.65		0.67	0.58
NY	Grade 7	0.55	0.60		0.66	0.59
NY	Grade 8	0.70	0.62		0.64	0.61
Texas	Grade 3		0.68	0.62	0.68	
Texas	Grade 4		0.66	0.62	0.70	
Texas	Grade 5		0.70	0.68	0.74	
Texas	Grade 6		0.64	0.64	0.65	
Texas	Grade 7		0.67	0.65	0.70	
Texas	Grade 8		0.70	0.67	0.71	

C NWEA Spring 2020 Reading Student Achievement Norms

Table 4: NWEA Spring 2020 Reading Student Achievement Norms by Grade

Grade	Score
Grade 3	200.74
Grade 4	204.83
Grade 5	210.19
Grade 6	215.36
Grade 7	216.81
Grade 8	220.93

D Passage and Item Characteristics Variables

Table 5: Variables: Passage and Item Characteristics

Variable	Description
item_order	Number indicating the order in which item appears after a passage
pass_highlight_yn	Indicator (1/0) if the passage text is highlighted
ques_text_ref_yn	Indicator (1/0) if the question text includes relevant paragraphs or sentences from text
ques_text_highlight_yn	Indicator (1/0) if the question text includes text in bold or underlined

E Variables Used as Predictors

E.1 Linguistic Features

Table 6: Table of Measures and Descriptions

Measure	Description
Descriptive metrics	
PassageWordCount	Word count of the passage
DESPC	Paragraph count, number of paragraphs
DESSC	Sentence count, number of sentences
DESPL	Paragraph length, number of sentences in a paragraph, mean
DESPLd	Paragraph length, number of sentences in a paragraph, standard deviation
DESSL	Sentence length, number of words, mean
DESSLd	Sentence length, number of words, standard deviation
DESWLsy	Word length, number of syllables, mean
DESWLsyd	Word length, number of syllables, standard deviation
DESWLlt	Word length, number of letters, mean
DESWLltd	Word length, number of letters, standard deviation
Readability metrics	
FK	Flesch-Kincaid score
RDFRE	Flesch-Kincaid Reading Ease
RDFKGL	Flesch-Kincaid Grade level
RDL2	Coh-Metrix L2 Readability
Text reading "easability": Principal Components analysis mapped to dimensions of reading ease i.e., syntactic simplicity, word concreteness, referential cohesion, deep cohesion, verb cohesion, connectivity, and temporality	
PCNARz	Text Easability PC Narrativity, z score (e.g., tells a story)
PCNARp	Text Easability PC Narrativity, percentile
PCSYNz	Text Easability PC Syntactic simplicity, z score (e.g., sentences contain fewer words and simpler sentences)
PCSYNp	Text Easability PC Syntactic simplicity, percentile
PCCNCz	Text Easability PC Word concreteness, z score (e.g., more concrete words)
PCCNCp	Text Easability PC Word concreteness, percentile
PCREFz	Text Easability PC Referential cohesion, z score (e.g., words or ideas repeat across text)
PCREFp	Text Easability PC Referential cohesion, percentile
PCDCz	Text Easability PC Deep cohesion, z score (e.g., text contains causal and intentional connectives that aid with comprehension)
PCDCp	Text Easability PC Deep cohesion, percentile
PCVERBz	Text Easability PC Verb cohesion, z score (e.g., overlapping verbs in text)
PCVERBp	Text Easability PC Verb cohesion, percentile
PCCONNz	Text Easability PC Connectivity, z score (e.g., contains explicit adversative, additive, and comparative connectives)
PCCONNp	Text Easability PC Connectivity, percentile
PCTEMPz	Text Easability PC Temporality, z score (e.g., text has cues about temporality such as tense and aspect)
PCTEMPp	Text Easability PC Temporality, percentile
Referential cohesion: Word overlap across text, measured by indices of word and sentence overlaps	
CRFNO1	Noun overlap, adjacent sentences, binary, mean
CRFAO1	Argument overlap, adjacent sentences, binary, mean
CRFSO1	Stem overlap, adjacent sentences, binary, mean

Measure	Description
CRFNOa	Noun overlap, all sentences, binary, mean
CRFAOa	Argument overlap, all sentences, binary, mean
CRFSOa	Stem overlap, all sentences, binary, mean
CRFCWO1	Content word overlap, adjacent sentences, proportional, mean
CRFCWO1d	Content word overlap, adjacent sentences, proportional, standard deviation
CRFCWOa	Content word overlap, all sentences, proportional, mean
CRFCWOad	Content word overlap, all sentences, proportional, standard deviation
Referential cohesion: Semantic overlap in text measured by Latent Semantic Analysis	
LSASS1	LSA overlap, adjacent sentences, mean
LSASS1d	LSA overlap, adjacent sentences, standard deviation
LSASSp	LSA overlap, all sentences in paragraph, mean
LSASSpd	LSA overlap, all sentences in paragraph, standard deviation
LSAPP1	LSA overlap, adjacent paragraphs, mean
LSAPP1d	LSA overlap, adjacent paragraphs, standard deviation
LSAGN	LSA given/new, sentences, mean
LSAGNd	LSA given/new, sentences, standard deviation
Lexical diversity: Variety of words in a given text. Lexical diversity is lower and cohesion is higher when words are repeated across text	
LDTTRc	Lexical diversity, type-token ratio, content word lemmas
LDTTRa	Lexical diversity, type-token ratio, all words
LDMTLD	Lexical diversity, MTLD, all words
LDVOCD	Lexical diversity, VOCD, all words
Connectives: Incidence of connectives per 1000 words, which reflects cohesive links between ideas and text organization. Cohesion is higher when there is greater incidence of connectives.	
CNCAll	All connectives incidence
CNCCaus	Causal connectives incidence (e.g., *because, so*)
CNCLogic	Logical connectives incidence (e.g., *and, or*)
CNCADC	Adversative and contrastive connectives incidence (e.g., *although, whereas*)
CNCTemp	Temporal connectives incidence (e.g., *first, until*)
CNCTempx	Expanded temporal connectives incidence
CNCAdd	Additive connectives incidence (e.g., *and, moreover*)
CNCPos	Positive connectives incidence (e.g., *also, moreover*)
CNCNeg	Negative connectives incidence
Situation model: Representations of deeper meaning in text using Lexical Semantic Analysis and WordNet	
SMCAUSv	Causal verb incidence (e.g., *although, whereas*)
SMCAUSvp	Causal verbs and causal particles incidence
SMINTEp	Intentional verbs incidence
SMCAUSr	Ratio of causal particles to causal verbs
SMINTEr	Ratio of intentional particles to intentional verbs
SMCAUSlsa	LSA verb overlap
SMCAUSwn	WordNet verb overlap
SMTEMP	Temporal cohesion, tense and aspect repetition, mean
Syntactic complexity: Includes measures of complex syntax in text and Syntactic Pattern Density; where higher density reflects more informational dense text with more complex syntax	
SYNLE	Left embeddedness, words before main verb, mean
SYNNP	Number of modifiers per noun phrase, mean
SYNMEDpos	Minimal Edit Distance, part of speech
SYNMEDwrd	Minimal Edit Distance, all words
SYNMEDlem	Minimal Edit Distance, lemmas

Measure	Description
SYNSTRUTa	Sentence syntax similarity, adjacent sentences, mean
SYNSTRUTt	Sentence syntax similarity, all combinations, across paragraphs, mean
DRNP	Noun phrase density, incidence
DRVP	Verb phrase density, incidence
DRAP	Adverbial phrase density, incidence
DRPP	Preposition phrase density, incidence
DRPVAL	Agentless passive voice density, incidence
DRNEG	Negation density, incidence
DRGERUND	Gerund density, incidence
DRINF	Infinitive density, incidence
Word complexity: information about parts of speech (e.g., relative frequency of types of word categories such as nouns, verbs and adverbs), word frequency (e.g., frequency of words used in text in CELEX database), and word complexity based on psychological ratings (e.g., age of acquisition of words, word concreteness)	
WRDNOUN	Noun incidence
WRDVERB	Verb incidence
WRDADJ	Adjective incidence
WRDADV	Adverb incidence
WRDPRO	Pronoun incidence
WRDPRP1s	First person singular pronoun incidence
WRDPRP1p	First person plural pronoun incidence
WRDPRP2	Second person pronoun incidence
WRDPRP3s	Third person singular pronoun incidence
WRDPRP3p	Third person plural pronoun incidence
WRDFRQc	Word frequency for content words using CELEX database, mean
WRDFRQa	CELEX Log frequency for all words, mean
WRDFRQmc	CELEX Log minimum frequency for content words, mean
WRDAOAc	Age of acquisition for content words, mean
WRDFAMc	Familiarity for content words, mean
WRDCNCc	Concreteness for content words, mean
WRDIMGc	Imagability for content words, mean
WRDMEAc	Meaningfulness, Colorado norms, content words, mean
WRDPOLc	Polysemy for content words, mean
WRDHYPn	Hypernymy for nouns, mean
WRDHYPv	Hypernymy for verbs, mean
WRDHYPnv	Hypernymy for nouns and verbs, mean

F Predictors for Item Difficulty Model

F.1 Visual representation of predictors in Item Difficulty Model

The figure shows a sample item from New York 2019 Grade 4 test. The box **"Input text for generating embeddings"** demonstrates how the passage and item is converted into a sentence. Embeddings are generated for this sentence. The boxes on the right show samples of annotated features for the item: **Linguistic features** shows some examples of text analysis of the reading passage. **Test features** show the relevant item design elements. **Context features** capture the item's year, state, and targeted grade.

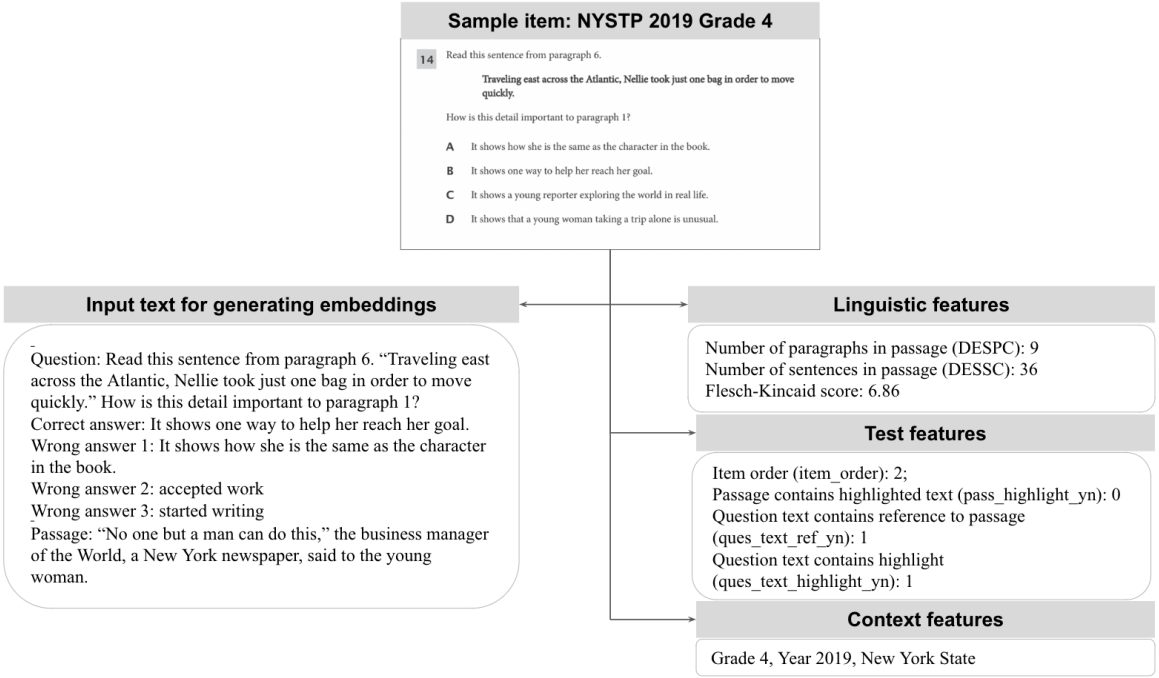

Figure 2: Predictors for Item Difficulty Model: Input Text to Generate Embeddings, and Linguistic, Test, and Context Features

F.2 Top predictors of Item Difficulty from Ridge Regression

Table 7: Top 20 Predictors of Item Difficulty from Ridge Regression (Text Analysis and Human-Annotated Features)

Rank	Variable	Coefficient	Absolute Value
Referential Cohesion			
1	CRFANP1 (Noun phrase overlap, adjacent sentences)	−0.127	0.127
14	LSAPP1 (LSA overlap, adjacent paragraphs)	−0.065	0.065
15	LSAGNd (LSA given/new, sentences SD)	+0.065	0.065
16	LSASS1 (LSA overlap, adjacent sentences)	+0.063	0.063
Syntactic Complexity			
2	DRNP (Noun phrase density)	−0.123	0.123
5	SYNNP (Modifiers per noun phrase)	−0.106	0.106
7	SYNMEDlem (Minimal edit distance, lemmas)	+0.104	0.104
19	DRGERUND (Gerund density)	+0.056	0.056
Descriptive Metrics			
3	DESSC (Sentence count)	+0.109	0.109
9	DESPC (Paragraph count)	−0.093	0.093
12	DESWC (Passage word count)	−0.069	0.069
20	PassageWordCount.gen (Passage word count, generated)	−0.056	0.056
Test Features (Human Annotated)			
4	ques_text_highlight_yn (Question text contains highlight)	+0.108	0.108
Connectives			
6	CNCNeg (Negative connectives incidence)	+0.104	0.104
10	CNCADC (Adversative & contrastive connectives)	−0.077	0.077
Word Complexity			
8	WRDFRQa (Word frequency, all words)	+0.097	0.097
11	WRDIMGc (Imageability, content words)	+0.070	0.070
13	WRDNOUN (Noun incidence)	+0.066	0.066
Situation Model			
17	SMCAUSwn (WordNet verb overlap)	+0.063	0.063
18	SMINTEp (Intentional verb incidence)	+0.061	0.061

Note. Coefficients are from a Ridge (L2) regression predicting IRT-scaled item easiness. All predictors are standardized (mean = 0, SD = 1). Positive coefficients indicate association with easier items; negative coefficients indicate association with harder items. Only text analysis and human-annotated features are shown; embedding predictors are excluded as their coefficients are not directly interpretable.

G Robustness Checks

G.1 Robustness to Choice of Grade-Level Growth Scale

This section examines the robustness of results to selecting an alternate vertical scale to convert p-values to IRT difficulty. Table 8 describes alternate scales that could be used: NWEA 2015 MAP reading achievement standard (reported separately for literary texts and informational text), NWEA 2020 reading achievement as measured in Fall or Winter, Texas STAAR 2023-24 performance standards ("approaches grade-level", "meets grade level", and "masters grade level" performance), and Texas STAAR 2017-18 readiness standard. These different scales were selected to account for differences in state contexts, as well as differences in grade level scales before and after 2020.

Results for test RMSE, as well as correlation between true and predicted RMSE are shown in Figure 3. We note that the results presented in the main section are conservative and do not fluctuate drastically with choice of scale. The greatest variation is observed when we use NWEA Informational Text 2015 scale for 2018 and 2019, and the NWEA 2020 Spring scale for 2021 to 2023. The RMSE increases to 3.24, likely due to the introduction of variation across years. However, correlation between true and predicted difficulty also increases to 0.96. Since it is difficult to obtain precise estimates of the change in vertical grade level scales before and after 2020, these results are not highlighted in the main paper.

Table 8: Alternate Vertical Growth Scales: NWEA MAP 2015, NWEA MAP 2020 and Texas Meets Grade Level Performance 2023-24

Grades	NWEA 2015		NWEA 2020			Texas STAAR 2023-24			Texas 2017-18
	Literary Text	Informational Text	Fall	Winter	Spring	Approaches grade-level	Meets grade-level	Masters grade-level	Readiness standard
Grade 3	192.4	191.6	186.62	195.91	200.74	1345	1467	1596	1386
Grade 4	201.2	200.7	196.67	202.5	204.83	1414	1552	1663	1473
Grade 5	207.9	207.4	204.48	210.19	210.98	1471	1592	1700	1508
Grade 6	212.3	212.1	210.17	213.81	215.36	1535	1634	1749	1554
Grade 7	216.3	216.1	214.2	217.09	216.81	1564	1669	1771	1603
Grade 8	220.0	220.0	218.9	220.52	220.93	1592	1698	1803	1625

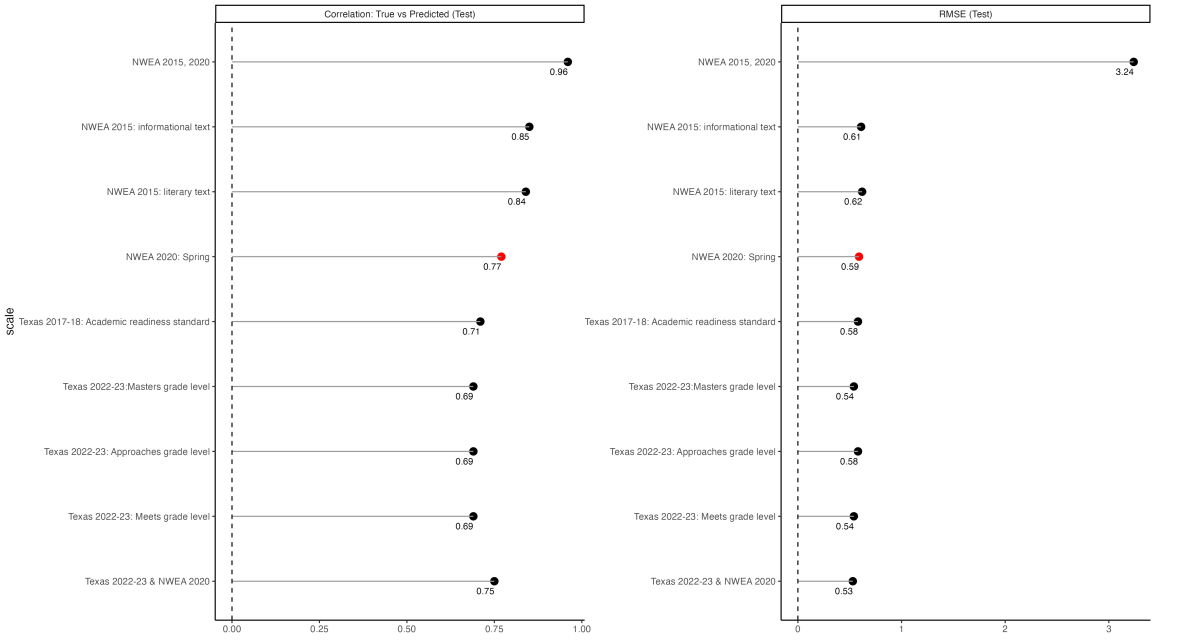

Figure 3: **Robustness to Vertical Scale Choice**

Note. (Left panel) Correlation of true vs predicted difficulty for test data; (Right panel) RMSE for true vs predicted difficulty. The main results reported in the paper use the NWEA Spring 2020 vertical scale, shown in red in the figure. Note that results reported in the paper are conservative and not extremely sensitive to the choice of vertical scale.

G.2 BERT Embeddings: Robustness to Text Input Format

Table 9: BERT Model: Robustness of Prediction Results

	RMSE		Correlation	
	Train	**Test**	**Train**	**Test**
Main Results: Text input is Question + Correct answer + Wrong answers + Passage				
LLM embeddings: BERT	0.60	0.66	0.76	0.71
Assessment characteristics, text analysis metrics, & BERT embeddings	0.60	0.64	0.76	0.73
Assessment characteristics, text analysis metrics, & PCA on BERT embeddings	0.58	0.64	0.76	0.72
Alternate Results 1: Text input is Question + Correct answer + Wrong answers				
LLM embeddings: BERT	0.65	0.72	0.71	0.63
Assessment characteristics, text analysis metrics, & BERT embeddings	0.60	0.65	0.76	0.73
Assessment characteristics, text analysis metrics, & PCA on BERT embeddings	0.58	0.61	0.76	0.75
Alternate Results 2: Main results with Cosine Similarity Variables				
LLM embeddings: BERT	0.60	0.66	0.76	0.72
Assessment characteristics, text analysis metrics, & BERT embeddings	0.60	0.64	0.76	0.74
Assessment characteristics, text analysis metrics, & PCA on BERT embeddings	0.58	0.63	0.76	0.73

The main results report BERT embeddings generated for a statement that merges question text, correct answer, all wrong answers, and passage (Appendix F). The reading passage was removed from the statement used to generate BERT embeddings for `Alternate Results 1` (Table 9). If the passage is relevant to how BERT embeddings capture information about the item, we would expect decline in model performance. We do observe this decline: RMSE increases from 0.66 for `Main Results` to 0.72 for `Alternate Results 1`, and correlation decreases from 0.72 to 0.63. However, prediction results are similar when human annotated features are included in the model, and when the model uses PCA on BERT embeddings instead of the full embeddings. This suggests that human annotated features might contain complementary information that compensates model performance. This is also interesting because BERT embeddings are generated using a section of the passage (sentence length is truncated at 512 characters). This suggests that a segment of passage still includes useful information; hence, the main results use embeddings generated with the passage included in the text input to the tokenizer.

We also include new predictor variables: cosine similarity is calculated between the embeddings for correct answer and each of the distractors. Embeddings are generated for the sentence that combines: (1) question text (2) a tag for correct/wrong answer, followed by the correct/wrong answer. Next, we calculate cosine similarity between the embeddings for the correct answer and each of the wrong answers, giving us three cosine similarity variables. These variables are included as predictors in the main model. These results are reported as `Alternate Results 2`. We note marginal improvement in performance between `Main Results` and `Alternate Results 2`, where RMSE decreases from 0.64 to 0.63, and correlation marginally improves from 0.72 to 0.73. This suggests that differences or similarities in BERT embeddings for correct answers and distractors don't explain item difficulty in the context of this dataset.

G.3 Results for Unadjusted p-values

This section presents the results for unadjusted p-values i.e., the p-values reported in the dataset that have not been adjusted for grade-level differences. As discussed in the methods section, this measure of difficulty does not have an inherent meaning. The results show a lower RMSE, likely because grade-level variation has not been accounted for. However, the correlation is also low (in the range of 0.10–0.3).

Table 10: Results for Predicting Item Difficulty: Outcome Variable is Unadjusted p-value

	RMSE		**Correlation**	
	Train	**Test**	**Train**	**Test**
Results from human annotated features				
State, Grade, Year	0.12	0.12	0.28	0.29
Test features	0.13	0.13	0.15	0.10
Text analysis features	0.12	0.13	0.36	0.11
Results from LLM embeddings only				
Only BERT embeddings	0.11	0.12	0.57	0.30
Only LLaMA embeddings	0.10	0.12	0.69	0.35
Results from LLM embeddings and annotated features				
All features & BERT embeddings	0.10	0.12	0.60	0.32
All features & PCA on BERT embeddings	0.12	0.12	0.46	0.25
All features & LLaMA embeddings	0.10	0.12	0.72	0.35
All features & PCA on LLaMA embeddings	0.12	0.12	0.46	0.25
All features & PCA on ModernBERT embeddings	0.12	0.12	0.45	0.25