

# Evaluating Multiple Models for Predicting Item Difficulty in a Principled Assessment Design Framework

Alexandra Perez, M. Christina Schneider, Sangdon Lim, Garron Gianopulos, and Kang Xue

Cambium Assessment, Inc.

## Abstract

Item difficulty should, in theory, be predictable from features derived from RPLDs, task characteristics, and linguistic complexity. This study evaluates the performance of multiple statistical and machine learning models, linear regression, LASSO regression, random forest, and gradient boosting, in estimating item difficulty. We found similar results across all four models, the item features selected explain 53%–56% of the variance across all grades. Across all four models, number of words in an item, Amount of Text/Source of Inference and grade were the top three important features in predicting item difficulty.

## 1 Background

Recent advances in assessment design emphasize the integration of cognitive theory, task models, and the use of construct relevant features intended to drive item difficulty (difficulty drivers) to engineer items in particular performance levels (Huff et al., 2025). Within this paradigm, principled assessment design (PAD) links both Range performance level descriptors (RPLDs; Egan et al., 2012) and difficulty drivers to task characteristics to support claims about what students know and can do.

RPLDs are intended to provide descriptions of increasing knowledge, skills, and abilities that are expected of students based on a state's content standards as student abilities increase along a test scale. They are intended to provide content-based interpretations regarding what

scale scores within a particular performance level represent within a grade. They are also intended to describe text-task interactions (Schneider et al., 2026) when developed for reading. Text-task interactions are a learning-science-based theory of reading comprehension (Valencia et al., 2017). RPLDs may have different characteristics based on test developers' theory of reading development; however, following the Egan et al. (2012) development framework, RPLDs for English language arts (ELA) optimally describe text-task interactions rather than text complexity. When validating assessments developed under a PAD framework, item difficulty modeling (IDM) is a critical component.

Item difficulty should, in theory, be predictable from features derived from RPLDs, task characteristics, and linguistic complexity. In the future, the goal is for these features to be transparent and useful in supporting item writers and their Large Language Model (LLM) assistants in constructing items with differing characteristics that align to specific areas of the test scale that represent changing stages of student cognition. Range PLDs have been growing in use as a component of item design. Researchers have increasingly applied machine learning techniques and LLMs to predict item difficulty.

Sano (2016) used a linear regression of stem, answer choice options, and passage textual complexity features of 60 Grade 8 NAEP items and found 9 of the 11 predictors that were significant related to NLP features parsed separately for stems, the key, and the options. He highlighted correlations of features that reached

.30. Štěpánek et al. (2023) found LASSO regression, ridge regression, and elastic net to perform the best, having the lowest root mean squared error (RMSE) when comparing models. Razavi and Powers (2025) used LLMs to derive item features to predict item difficulty. They conducted an IDM study and found that the random forest model and gradient boosting model produced similar results.

Machine learning techniques have included linear regression (Hastie et al., 2009), random forests (Breiman, 2001), LASSO regression (Tibshirani, 1996), and gradient boosting (XGBoost; Chen & Guestrin, 2016). However, we believe there remains a need to evaluate how different modeling approaches perform within a PAD/IDM framework which values feature transparency to determine which features most strongly predict item difficulty across grade levels in construct relevant ways.

This study investigates the extent to which item characteristics derived from a combination of human coded features and feature extractions predict item difficulty. ELA items have historically been difficult to predict (Ferrara et al., 2022) with  $R^2$  above .50 in a student population. Our aims follow:

1. Evaluate the predictive performance of multiple statistical and machine learning models, including linear regression, random forest, LASSO regression, and gradient boosting (XGBoost), in estimating item difficulty.
2. Identify the most influential item features contributing to difficulty across grades 3–8 in an ELA state assessment.
3. Inform item development practices, with the goal of guiding item writers toward features that predict intended difficulty levels.

## 2 Data

The data for this study came from items and human coding of those items measuring ELA standards for grades 3–8. Psychometricians used the 2-parameter model for dichotomous items and the generalized partial credit model (GPCM; Muraki, 1992) for polytomous items for vertical scaling. Items were located at the point of the scale where the response probability was .50 (RP50) for the score point or maximum score point for polytomous items. There were 1,264 items, and the number of items per grade ranged from 128 to 288, with grades 4 and 7 having the fewest items.

## 3 Methods

The task features came from the item banking software, human derived features, the Python Natural Language Toolkit (NLTK; Bird et al., 2009), and Python's textstat library (Shivam, 2024).

To determine the set of predictors included in our model, we used principal components analysis (PCA). However, because many variables make up these components, interpreting them and using the results in practice can be challenging. Therefore, we extracted the components that contributed to 95% of the variance and selected the variable that had the highest contribution in each component. This allowed us to be more confident that the variable importance measures were not influenced by correlated predictors. The final set of predictors included item RPLD level, Amount of Text/Source of Inference (Valencia et al., 2017), Grade, binary item types (multiple select, other, and EBSR), and item text features (number of words, Linsear Write index, Dale Chall index, Coleman Liau index, and whether the response options included a 'paragraph prompt').

We employed separate model-training and model-testing processes for the models, ensuring a more accurate evaluation using a 5-fold cross validation. We split the data into five folds (parts). Each model was trained on four folds and tested on the remaining fold. Because this was repeated five times, each fold was used as the test set once. For model tuning, we repeated this 5-fold cross validation 10 times. For model analysis, we repeated this 5-fold cross validation 100 times.

We compare Ordinary Least Squares (OLS) linear regression, LASSO regression, random forest regression model, and gradient boosting (XGBoost) model. To compare model performance, we examined RMSE and  $R^2$ .

For LASSO regression, we initially tuned the model to find the best lambda, as lambda approaches zero, the closer the model is to linear regression (Tibshirani, 1996). For the random forest model, we initially tuned the model to find the best combination of the following hyperparameters:

- number of trees,
- maximum number of splits from the root to the leaf (max depth),
- number of features randomly considered at each split (mtry),
- and the smallest number of observations allowed in a leaf (minimum node size).

Finally, for the gradient boosting model, we initially tuned the model to find the best combination of the following hyperparameters:

- how much each new tree is allowed to adjust the model's predictions (eta),
- how complex each tree can be (max depth),

- how much data is required before a split is allowed (minimum child weight),
- the fraction of rows used for each tree (subsample),
- and the fraction of predictors used for each tree (column sample by tree).

## 4 Results

Table 1 and Table 2 show the results from the 5-fold cross validation of the linear regression model, the LASSO regression model, the random forest regression model, and the gradient boosting model. For the individual grade-level analyses, we used the predictions from the testing data and calculated  $R^2$  separately for each grade by subsetting the results by grade level. Linear regression and LASSO regression expect linear relationships whereas random forests and gradient boosting show the influence of interactions of features.

The amount of variance explained and RMSE in the testing data for all four models were similar across all grades and within grades. The item features selected explain 53%–56% of the variance across all grades. The RMSE is approximately 0.8. Additionally, for the  $R^2$  and RMSE, the difference between the testing and training dataset was not large, therefore there is minimal overfitting. The random forest model revealed the greatest degree of overfitting, as evidenced by the largest discrepancy between training and testing performance. The lambda value for the LASSO regression model was 0.004, indicating that this model performed similarly to the linear regression model.

**Table 1. Average Variance Explained ( $R^2$ ) Comparative Results**

| Average $R^2$ of Training and Testing Data |               |              |         |         |         |         |         |         |
|--------------------------------------------|---------------|--------------|---------|---------|---------|---------|---------|---------|
|                                            | Training Data | Testing Data |         |         |         |         |         |         |
|                                            | All Grades    | All Grades   | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 |
| Linear Regression                          | 0.55          | 0.53         | 0.53    | 0.09    | 0.58    | 0.38    | 0.33    | 0.37    |
| LASSO Regression                           | 0.55          | 0.53         | 0.53    | 0.10    | 0.58    | 0.38    | 0.33    | 0.37    |
| Random Forest                              | 0.71          | 0.55         | 0.59    | 0.12    | 0.61    | 0.36    | 0.27    | 0.40    |
| Gradient Boosting                          | 0.67          | 0.56         | 0.59    | 0.11    | 0.60    | 0.37    | 0.31    | 0.41    |

*Note.* The outcome was RP50 of items. We employed a 5-fold CV process and repeated it 100 times.

**Table 2. Average Root Mean Squared Error (RMSE) Comparative Results**

| Average RMSE of Training and Testing Data |               |              |         |         |         |         |         |         |
|-------------------------------------------|---------------|--------------|---------|---------|---------|---------|---------|---------|
|                                           | Training Data | Testing Data |         |         |         |         |         |         |
|                                           | All Grades    | All Grades   | Grade 3 | Grade 4 | Grade 5 | Grade 6 | Grade 7 | Grade 8 |
| Linear Regression                         | 0.84          | 0.85         | 0.95    | 0.83    | 0.82    | 0.74    | 0.89    | 0.82    |
| LASSO Regression                          | 0.84          | 0.85         | 0.95    | 0.83    | 0.82    | 0.74    | 0.89    | 0.82    |
| Random Forest                             | 0.67          | 0.83         | 0.88    | 0.82    | 0.78    | 0.76    | 0.93    | 0.80    |
| Gradient Boosting                         | 0.71          | 0.83         | 0.87    | 0.82    | 0.79    | 0.75    | 0.91    | 0.80    |

*Note.* The outcome was RP50 of items. We employed a 5-fold CV process and repeated it 100 times.

The results were not consistent across grades. The variance explained was strongest for Grades 3 and 5 and slightly lower for Grades 6 through 8. Grade 4 was an outlier in comparison to other grades. Grade 4 could be lower because the construct differs, because of the smaller sample size, or a combination. Grades 4 and 7 had lower percentages of items aligned to RPLD Level 4 and Text Source Level 4 compared to other grades.

Table 3 presents evaluation results from the linear regression model. These results represent averages across the 100 5-fold cross-validation datasets across grades. The significant features are highlighted in both models. The results are

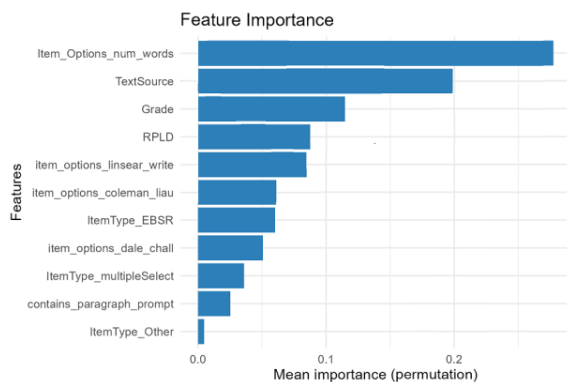
similar across the two models. All variables are statistically significant except for the item containing a paragraph prompt, Dale Chall index, and Coleman Liau Index. The significant predictors indicate that as the predictor increases, the item difficulty increases, holding all else constant. If an item is an EBSR item, multiple select item, or other item type, the item difficulty increases indicating that multiple choice items predict lower item difficulty for these items. For LASSO regression, similar features were found important in predicting item difficulty, number of words in an item, Amount of Text/Source of Inference, Grade, and the three item type features.

**Table 3. Average Regression Coefficients over Replications Across Grades**

|                           | <i>b</i> | <i>p</i> | CI low | CI high |
|---------------------------|----------|----------|--------|---------|
| (Intercept)               | -3.21    | 0.00     | -3.59  | -2.83   |
| # of Words                | 0.01     | 0.00     | 0.01   | 0.01    |
| Linsear Write Index       | 0.03     | 0.00     | 0.01   | 0.04    |
| Contains Paragraph Prompt | -0.04    | 0.53     | -0.17  | 0.09    |
| Item Type – MS            | 0.82     | 0.00     | 0.63   | 1.01    |
| Item Type -Other          | 0.30     | 0.01     | 0.09   | 0.50    |
| Amount of Text Source     | 0.32     | 0.00     | 0.25   | 0.38    |
| Item Type – EBSR          | 0.44     | 0.00     | 0.25   | 0.64    |
| Dale Chall Index          | 0.05     | 0.08     | 0.00   | 0.10    |
| RPLD                      | 0.07     | 0.04     | 0.01   | 0.14    |
| Coleman Liau Index        | 0.00     | 0.72     | -0.03  | 0.03    |
| Grade                     | 0.16     | 0.00     | 0.12   | 0.19    |

*Note.* The outcome was RP50 of items. We employed a 5-fold CV process and repeated it 100 times. We averaged the coefficients across the training datasets.

Figure 1 shows the feature importance of the random forest model. As shown, number of words, Amount of Text/Source of Inference, Grade, and RPLD are the most important features predicting item difficulty in order of performance. For the gradient boosting model, number of words, Amount of Text/Source of Inference and Grade were the strongest features.

**Figure 1: Feature Importance from Random Forest Model**

Across all four models, the most important features predicting item difficulty were number of words, Amount of Text/Source of Inference and Grade. For linear regression and LASSO regression, we standardized the coefficients to ensure fair comparison to random forest and gradient boosting. While RPLD is a significant feature in the models, it does not predict an increase in item difficulty as great as Amount of Text/Source of Inference. See Table 4 for the list of important features by model. These features may be contextualized into features related to stem and options reading load and vocabulary, source of evidence in the text needed to respond correctly, RPLDs, and item types.

**Table 4. Feature Importance Ordered by Model**

| Importance | Linear Regression                  | LASSO Regression                   | Random Forest                      | XGBoost                            |
|------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| 1          | Number of Words                    | Number of Words                    | Number of Words                    | Number of Words                    |
| 2          | Amount of Text/Source of Inference | Amount of Text/Source of Inference | Amount of Text/Source of Inference | Amount of Text/Source of Inference |
| 3          | Grade                              | Grade                              | Grade                              | Grade                              |
| 4          | Multiple Select Item Type          | Multiple Select Item Type          | RPLD                               | Dale Chall Index                   |
| 5          | EBSR Item Type                     | EBSR Item Type                     | Linsear Write Index                | Multiple Select Item Type          |
| 6          | Linsear Write Index                | Linsear Write Index                | Coleman Liau Index                 | Coleman Liau Index                 |
| 7          | Other Item Type                    | RPLD                               | EBSR Item Type                     | Linsear Write Index                |
| 8          | Dale Chall Index                   | Dale Chall Index                   | Dale Chall Index                   | RPLD                               |
| 9          | RPLD                               | Other Item Type                    | Multiple Select Item Type          | EBSR Item Type                     |
| 10         | Coleman Liau Index                 | Contains Paragraph Prompt          | Contains Paragraph Prompt          | Other Item Type                    |
| 11         | Contains Paragraph Prompt          | Coleman Liau Index                 | Other Item Type                    | Contains Paragraph Prompt          |

## 5 Conclusion and Discussion

Overall, the variables we included in the model accounted for more than 50% of the variance across all grades. There was minimal overfitting in the models as the training dataset's average  $R^2$  and RMSE were similar for the testing dataset, specifically for the linear regression and LASSO regression. As expected, random forest and gradient boosting had more overfitting, but gradient boosting outperformed random forest in this aspect. Within grades, the average variance explained dropped in grades 4, 6, 7, and 8. For these grades, we could have seen a drop in performance if the construct differs across grades; that is, if the feature combinations were being deployed differently across grades.

Across all four models, number of words in an item, Amount of Text/Source of Inference and grade were the top three important features in predicting item difficulty. For linear regression and LASSO regression, the item type features also contributed to the variation in item difficulty. For the random forest model, the next most important feature was RPLD.

For this study, we found that the four models performed similarly to each other, but the

features explaining the variation in item difficulty could differ. If item writers need to craft items that are likely to be more difficult, they should find opportunities to ask questions using text characteristics at the higher end of the Amount of Text/Source of Inference rubric. These types of questions also will tend to involve longer answer choice options and target the higher end of the RPLDs. However, based on the inconsistent finding of RPLDs across models, we recommend that Amount of Text/Source of Inference be a difficulty driver embedded in the RPLDs in the future.

## 6 References

- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.  
<https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

- Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance-level descriptors. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.
- Ferrara, S., Steedle, J., & Franz, R. (2022). Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education*, 35(3), 237–253.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- Huff, K., Nichols, P., & Schneider, M. C. (2025). Designing and developing educational assessments. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 441–511). National Council on Measurement in Education. DOI: 10.1093/oso/9780197654965.003.0007
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>
- Razavi, P., & Powers, S. J. (2025). Estimating Item Difficulty Using Large Language Models and Tree-Based Machine Learning Algorithms. *10.48550/arXiv.2504.08804*.
- Sano, M. (2016 April). Improvements in automated capturing of psycho-linguistic features in reading assessment text. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.
- Schneider, M.C., Chen, J., & Heneger, J. (2026). Text complexity versus task complexity: Item difficulty modeling for reading items. *Practical Assessment, Research, & Evaluation*, 31(1)(5). <https://doi.org/10.7275/pare.2928>
- Shivam, S. (2024). textstat: A Python package for text statistics (Version 0.7.4) [Computer software]. <https://pypi.org/project/textstat/>.
- 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. <https://doi.org/10.3390/math11194104>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Valencia, S. W., Wixson, K. K., Ackerman, T., & Sanders, E. (2017). Identifying Text-Task-Reader Interactions Related to Item and Block Difficulty in the National Assessment for Educational Progress Reading Assessment.