Item Difficulty and Response Time Prediction with Large Language Models: An Empirical Analysis of USMLE Items

Okan Bulut, Guher Gorgun, Bin Tan
Measurement, Evaluation, and Data Science
Faculty of Education, University of Alberta, Canada
{bulut, gorgun, btan4}@ualberta.ca

Abstract

This paper summarizes our methodology and results for the BEA 2024 Shared Task. This competition focused on predicting item difficulty and response time for retired multiple-choice items from the United States Medical Licensing Examination® (USMLE®). We extracted linguistic features from the item stem and response options using multiple methods, including the BiomedBERT model, FastText embeddings, and Coh-Metrix. The extracted features were combined with additional features available in item metadata (e.g., item type) to predict item difficulty and average response time. The results showed that the BiomedBERT model was the most effective in predicting item difficulty, while the fine-tuned model based on FastText word embeddings was the best model for predicting response time.

1 Introduction

In standardized exams, the examination of item characteristics is highly crucial for ensuring the fairness and validity of test results. For example, the difficulty of items pertains to the likelihood of an examinee answering the items correctly. Incorporating a broad range of item difficulty levels in a standardized exam can help reduce measurement error and thereby improve the accuracy of the measurement process (Kubiszyn and Borich, 2024). In addition, while response time is often linked to item difficulty (i.e., more difficult items require more time to answer) (Yang et al., 2002), this variable itself can also offer new insights into examinees' test completion processes, such as their testing engagement and cognitive processes, thereby supporting the validity of test results. Furthermore, understanding item characteristics can also be advantageous for modern test administration methods, including applications in automated item assembly, computerized adaptive testing, and personalized assessments (Baylari and Montazer, 2009; Wauters et al., 2012).

The difficulty of items and the average response time required to answer them are typically estimated based on empirical data collected during test pretesting. However, pretesting and obtaining robust results often require a large sample of examinees, which can incur substantial test administration costs. As a result, researchers have explored various methods to predict item characteristics without an actual test administration. For instance, researchers have sought estimates of item difficulty from domain experts and test development professionals. However, this approach has not consistently produced satisfactory or reliable estimations (Bejar, 1983; Attali et al., 2014; Wauters et al., 2012; Impara and Plake, 1998). Another line of research seeks to predict item characteristics based on only item texts, such as the passages in source-based items, item stem, and response options (Yaneva et al., 2019; Hsu et al., 2018). This approach employs text-mining techniques to extract surface features (e.g., the number of words in the texts) and complex features (e.g., semantic similarities of sentences) from item texts, to make predictions using advanced statistical models.

Building on the second line of research in predicting item characteristics based on item texts, the National Board of Medical Examiners (NBME) initiated the BEA 2024 Shared Task (https://sig-edu.org/sharedtask/2024) for automated prediction of item difficulty and item response time. The released dataset contained 667 previously used and now retired items from the United States Medical Licensing Examination® (USMLE®). The USMLE is a series of high-stakes examinations (also known as Steps; https://www.usmle.org/step-exams) to support medical licensure decisions in the United States. The items from USMLE Steps 1, 2 Clinical Knowledge (CK), and 3 focus on a wide range of topics relevant to the practice of medicine.

In the BEA 2024 Shared Task, research teams

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 522–527

were invited to utilize natural language processing (NLP) methods for extracting linguistic features of the items and using them to predict the difficulty and response time of the items. Our team employed state-of-the-art large language models (LLMs) to extract the features and build predictive models for item difficulty and response time. This paper documents the methods and results of our best-performing models for predicting item difficulty and response time separately.

2 Related work

The interest and effort in predicting item difficulty based on item texts dates back decades in the measurement literature. Early work in item difficulty prediction primarily focused on identifying how item difficulty is influenced by a set of readily available, easily extracted, or manually coded item-level features. For example, Drum et al. (1981) predicted the difficulty of 210 reading comprehension items using various surface structure variables and word frequency measures for the text, such as the number of words, content words, or content-function words. Freedle and Kostin (1993) predicted the difficulty of 213 reading comprehension items using 12 categories of sentential and discourse variables, such as vocabulary, length of texts, and syntactic structures (e.g., the number of negations). Perkins et al. (1995) employed artificial neural networks to predict the item difficulty of 29 items in a reading comprehension test. They coded the items to extract three types of features: text structure (e.g., the number of words, lines, paragraphs, sentences, and content words), propositional analysis of passages and stems (e.g., the number of arguments, modifiers, and predicates), and cognitive process (e.g., identify, recognize, verify, infer, generalize, or problem-solving).

Research focused on the prediction of item characteristics such as difficulty and response time has been significantly influenced by the availability and application of emerging techniques in NLP and machine learning AlKhuzaey et al. (2023). For example, Yaneva et al. (2019) employed NLP methods to extract syntactic features to predict item difficulty, which were identified as crucial predictors. Another application of NLP methods involves assessing the linguistic complexity or readability of item texts to predict item difficulty. Benedetto et al. (2020a), for instance, calculated readability indices for item texts and combined them with other features to predict item difficulty. However, readability indices did not perform well as predictors of item difficulty–a finding consistent with Susanti et al. (2017) who noted that readability indices were among the least important predictors of item difficulty.

NLP methods can also be used to extract Term Frequency-Inverse Document Frequency (TF-IDF) features. TF-IDF measures the frequency of words or word sequences in a document and adjusts this count based on their frequency across a collection of documents. This approach emphasizes the importance of specific words to a particular document, with higher values indicating greater potential importance (Salton, 1983). In a relatively recent study predicting item difficulty for newly generated multiple-choice questions, Benedetto et al. (2020b) extracted TF-IDF features and achieved a root mean square error of 0.753.

An important application of NLP techniques is the extraction of semantic features from item texts. Word embedding is a technique that converts texts into numerical values in vector space, capturing the meanings of words across different dimensions (Mikolov et al., 2013). Pre-trained NLP models such as Word2Vec and GLoVe allow researchers to extract word embedding features from item texts (e.g., Firoozi et al., 2022). For example, Hsu et al. (2018) transformed item texts into semantic vectors and then used cosine similarity to measure the semantic similarity between different pairs of items. Additionally, (Yaneva et al., 2019) extracted word embedding features from multiple-choice items in high-stakes medical exams. Along with other linguistic and psycholinguistic features in predicting item difficulty, they found that word embedding features contributed most to the predictive power.

More recently, a significant breakthrough in the NLP field has been the development of LLMs such as BERT (Devlin et al., 2018) and its variants, which were trained using different mechanisms or training datasets. For example, Zhou and Tao (2020) utilized a BERT-variant model to predict the difficulty of programming problems. Their results showed that compared with BERT, DistilBERT, a small version of the BERT base model, was the best-performing model when the only available data for fine-turning was the text of the items. Benedetto et al. (2021) also compared the performance of BERT and DistilBERT in predicting the difficulty of multiple-choice questions and found that the BERT-based models significantly outper-

formed the two baseline models.

Unlike the prediction of item difficulty, the prediction of response time has not been widely investigated in the literature. This is mainly due to the limited availability of response time data. However, with the increasing use of digital assessments, such as computer-based and computerized-adaptive tests, in operational testing, the collection of response data has become easier, which motivated researchers to employ predictive models to predict the average response time required to solve the items (e.g., Baldwin et al., 2021; Hecht et al., 2017; Yaneva et al., 2019).

3 Methodology

3.1 Dataset

As mentioned earlier, this study utilized an empirical dataset released by NBME, which included 667 multiple-choice items previously administered in the USMLE series. Due to the requirements of the BEA 2024 Shared Task, the data was released in two stages. Initially, 446 multiple-choice items were provided for extracting linguistic features from the items and building predictive models based on the extracted features. For each item, the dataset encompasses the source texts (typically a clinical case followed by a question) and the texts for each response option. The response options for the questions vary and can include up to 10 options, each represented in a separate column. When the number of response options was less than 10, the remaining columns were left empty.

Additionally, the dataset contained metadata with four additional variables: Item type (text-only items versus items containing pictures), exam steps (Steps 1, 2, or 3 in the USMLE series), item difficulty, and average response time. Subsequently, the predictive models trained in the first stage were applied to make predictions for the remaining 201 items in the second stage, serving as the testing set for evaluating the performance of the predictive models for item difficulty and response time. The structure of the second dataset mirrors that of the first, with the exception that the item difficulty and response time variables were not immediately available. These variables were released after the submission deadline for the BEA 2024 Shared Task, allowing for the identification of the best-performing trials among the participating teams.

3.2 Our Best Model for Difficulty Prediction

Here, we describe the details of our best-performing model for predicting item difficulty ($RMSE = .318$), which performed slightly worse than a baseline dummy regressor ($RMSE = .311$) and ranked at the 20th place out of 43 submissions in the difficulty prediction leaderboard.

3.2.1 Feature Extraction

We extracted linguistic features from item stems and response alternatives (i.e., the answer key and the incorrect response options) by leveraging both pre-trained large-language models and more interpretable text representations such as connectivity, cohesion, and text length. We started the feature extraction process by concatenating the stem, key, and alternatives of each item in a single data frame column and separated each item into individual data files to extract Coh-Metrix features (McNamara et al., 2014; Graesser et al., 2011). Concatenating item stems and alternatives served two purposes: (1) Adequately represent item length in terms of stem and alternatives and (2) control for the differential number of alternatives that each item includes. Coh-Metrix includes 108 features and analyzes a text on multiple measures of language and discourse (Graesser et al., 2011).

Coh-Metrix focuses on six theoretical levels of text representation: words, syntax, the explicit textbase, the referential situation model, the discourse genre and rhetorical structure, and the pragmatic communication level (Graesser et al., 2014). It generates indices of text, including paragraph count, sentence count, word count, narrativity, syntactic simplicity, referential cohesion, deep cohesion, noun overlap, stem overlap, latent semantic analysis, lexical diversity, syntactic complexity, syntactic pattern density, and readability. We removed four features from Coh-Metrix indices due to no variability, including paragraph count (i.e., the number of paragraphs), the standard deviation of paragraph length, the mean Latent Semantic Analysis (LSA) overlap in adjacent paragraphs, and the standard deviation of LSA overlap in adjacent paragraphs.

In the next step, we utilized the BiomedBERT model (Gu et al., 2020) to extract new features. This model, which was previously named PubMedBERT, is a pretrained LLM based on abstracts from PubMed and full-text articles from PubMedCentral. We chose this particular model because it is known to achieve state-of-the-art performance on

many biomedical NLP tasks. By using BiomedBERT, we obtained sentence embeddings for the item stems and alternatives and then computed the cosine similarity between item sentence embeddings and alternative stem embeddings. Cosine similarity, which is commonly used to quantify the degree of similarity between two sets of information, was computed as the cosine angle between the embedding vectors of item stem and alternatives. As cosine similarity, ranging between 0 and 1, gets closer to 1, it indicates more resemblance between the embedding vectors obtained using the item stem and alternatives.

In the final step, we also extracted word embeddings for the concatenated text using stems and alternatives by tokenizing the text using the BiomedBERT model (Gu et al., 2020). BiomedBERT has 768 dimensions with a maximum length of 512 words. We extracted the last hidden layer of embeddings. We created a new data frame composed of three sets of features extracted (i.e., Coh-Metrix features using the stem, key, and alternatives of each item, the cosine similarity between the stem and alternatives, and word embeddings using the stem, key, and alternatives) and the ground truth of item difficulty. The final data frame is composed of 882 features and the target variable of item difficulty.

3.2.2 Model Training

To identify the best model with the lowest RMSE value, we used 85% of the data as our training set and 15% as our holdout test set. Because the sample size was too small ($N = 466$ of items shared in total) and we had a very large set of features ($N = 882$), we first applied a dimension reduction technique, *Principal Component Analysis* (PCA) (Wold et al., 1987). A PCA model with 30 components explained 99% of variability in the dataset, and thus, the final feature set included 30 components extracted through the PCA analysis. We used lasso regression (Tibshirani, 1996) with repeated 5-fold cross-validation to select the best hyperparameter (i.e., *alpha*). *Alpha* in lasso regression is the model penalty that determines the amount of shrinkage in the model. An advantage of lasso regression is the application of a regularization algorithm that controls for the irrelevant features in the model by shrinking the contribution of irrelevant features to zero. An alpha value of .01 yielded the best model during the cross-validation stage.

3.2.3 Results

With our pseudo-test set held out from the shared training set, we obtained a Mean Squared Error (MSE) value of .064, a Root Mean Squared Error (RMSE) value of .253, and a Mean Absolute Error (MAE) value of .190, and a Pearson's correlation coefficient of .555.

3.3 Our Best Model for Response Time Prediction

Our solution that achieved the best performance in predicting response time differed from the one that was best at predicting item difficulty. This solution is briefly documented below.

3.3.1 Feature Extraction

First, FastText word embeddings were generated for each item stem and response option. We employed the pre-trained FastText embeddings (wiki-news-300d-1m.vec.zip; obtained from https://fasttext.cc/docs/en/english-vectors.html) to map each word in the text to its corresponding 300-dimensional vector representation. FastText is a modified version of word2vec; the difference is that it treats each word as composed of n-grams rather than the original word in Word2Vec (Mikolov et al., 2017). For each text option, the embeddings of the first 60 words were concatenated to form a feature vector, resulting in a dimension of 18,000 (60 words × 300 dimensions) for each option. If the text had fewer than 60 words, the corresponding vector was padded with zeros.

Similar to the approach taken for item difficulty predictions, cosine similarity scores between each pair of alternatives (i.e., response options) were calculated using the embeddings from the BiomedBERT model. For each pair, the cosine similarity between their embeddings was computed to capture the semantic differences between different response options. The extracted features were then combined with the dummy-coded item development information (e.g., text-based items only vs. items including pictures; administration step in the USMLE series) to form the final feature set. Unlike in the item difficulty prediction, we did not extract any other linguistic features in response time prediction.

3.3.2 Model Training

Considering the extremely high dimensionality of the features, we performed feature selection and dimension reduction techniques. First, using the

available information on response time in the training set ($N = 466$), we eliminated the feature columns that had an absolute correlation coefficient smaller than 0.1. Then, we performed PCA to extract principal components until they could capture 95% of the information presented in the original feature set. To this end, we obtained a final feature set with 339 features to train an algorithm.

As before, the model training involved the use of lasso regression due to its ability to perform feature selection and handle multicollinearity in high-dimensional data. The training process was performed using 10-fold cross-validation to optimize the hyperparameter (i.e., alpha) and evaluate the model's performance. In terms of the hyperparameter search space, the regularization strength (*alpha*) was tuned using a randomized search over a logarithmic scale from 1e-4 to 1e-0.05, with 1000 candidate values. An alpha value of .44 yielded the best model during the cross-validation stage. Additionally, the *fit intercept* parameter was tested with both True and False values, while the *selection parameter* was tested with 'cyclic' and 'random' options[1].

3.3.3 Results

Upon comparing our predicted response time and the released response time from the BEA 2024 Shared Task, we found this solution ($RMSE = 31.48$; $MSE = 990.98$; $MAE = 23.54$, $r = 0.209$) was slightly better than the baseline dummy regressor ($RMSE = 31.68$), which ranked 24$^{\text{th}}$ among the 34 submissions.

4 Discussion and Conclusion

The competition results for the BEA 2024 Shared Task indicated that it is difficult to predict item characteristics such as difficulty using linguistic features (Yaneva et al., 2024). Only 15 teams out of 43 managed to perform better than a baseline dummy regressor when it comes to predicting item difficulty using textual features extracted from the items. These results suggest that linguistic features may not be sufficient to capture the complex interplay between item features and item difficulty.

Unlike predicting item difficulty, predicting the average response time using linguistic features appears to be a more promising task. Out of 34 submissions, 24 teams performed better than a baseline dummy regressor in predicting the average response time. This finding is not necessarily surprising because the average reading time required for the items is likely to be correlated with the linguistic features extracted from the items.

Overall, the results for the BEA 2024 Shared Task indicate that predicting item characteristics such as difficulty remains challenging and requires factors beyond linguistic or textual features.

[1]Our codes for predicting item difficulty and response time are available at https://osf.io/dwe4n/.

References

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

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Ahmad Baylari and Gh A Montazer. 2009. Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications*, 36(4):8013–8021.

Isaac I Bejar. 1983. Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7(3):303–310.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th workshop on innovative use of NLP for building educational applications*, pages 147–157.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020a. Introducing a framework to assess newly created questions with natural language processing. In *International Conference on Artificial Intelligence in Education*, pages 43–54. Springer.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020b. R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 412–421.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Priscilla A Drum, Robert C Calfee, and Linda K Cook. 1981. The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, pages 486–514.

T Firoozi, O Bulut, C Demmans Epp, A N Abadi, and D Barbosa. 2022. The effect of fine-tuned word embedding techniques on the accuracy of automated essay scoring systems using neural networks. *Journal of Applied Testing Technology*, 23:21–29.

Roy Freedle and Irene Kostin. 1993. The prediction of toefl reading item difficulty: Implications for construct validity. *Language Testing*, 10(2):133–170.

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.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Martin Hecht, Thilo Siegle, and Sebastian Weirich. 2017. A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. *Journal for educational research online*, 9(1):32–51.

Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.

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.

Tom Kubiszyn and Gary D Borich. 2024. *Educational testing and measurement*. John Wiley & Sons.

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.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Kyle Perkins, Lalit Gupta, and Ravi Tammana. 1995. Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language testing*, 12(1):34–53.

Gerard Salton. 1983. Introduction to modern information retrieval. *McGraw-Hill*.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Controlling item difficulty for automatic vocabulary question generation. *Research and practice in technology enhanced learning*, 12:1–16.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 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.

Victoria Yaneva, Kai North, Peter Baldwin, An Ha Le, 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 (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Chien-Lin Yang, Thomas R O Neill, and Gene A Kramer. 2002. Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement*, 3(3):282–299.

Ya Zhou and Can Tao. 2020. Multi-task bert for problem difficulty prediction. In *2020 international conference on communications, information system and computer engineering (cisce)*, pages 213–216. IEEE.