Using Machine Learning to Predict Item Difficulty and Response Time in Medical Tests

Mehrdad Yousefpoori-Naeim
University of Alberta / Canada
yousefpo@ualberta.ca

Shayan Zargari
University of Alberta / Canada
zargari@ualberta.ca

Zahra Hatami
Epita Engineering School / Canada
zahra.hatami@epita.fr

Abstract

Prior knowledge of item characteristics, such as difficulty and response time, without pretesting items can substantially save time and cost in high-standard test development. Using a variety of machine learning (ML) algorithms, the present study explored several (non-)linguistic features (such as Coh-Metrix indices) along with MPNet word embeddings to predict the difficulty and response time of a sample of medical test items. In both prediction tasks, the contribution of embeddings to models already containing other features was found to be extremely limited. Moreover, a comparison of feature importance scores across the two prediction tasks revealed that cohesion-based features were the strongest predictors of difficulty, while the prediction of response time was primarily dependent on length-related features.

keywords: item difficulty, response time, machine learning, Coh-Metrix, MPNet embeddings

1 Introduction

Item difficulty and response time are among the important requirements in high-standard test development. For instance, in large-scale assessment, there is often a need to develop equivalent versions of the same test to be administered to different groups of people (DePascale and Gong, 2020). When deciding on the inclusion of items in each version, it is necessary to know the difficulty level of each item and an estimate of the time needed to answer that item. Such information is traditionally gained only through pretesting (Martinková and Hladká, 2023). Pretesting, however, is not a very efficient method, as it is expensive (Antal, 2013) and raises security concerns (Settles et al., 2020). Therefore, it would be highly advantageous to devise a method to ascertain item difficulty and response time without resorting to the pretesting of items.

With this motivation, a shared task was organized as part of the Building Educational Applications (BEA) workshop at the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2024. The shared task invited people to develop ML models for the prediction of item difficulty and response time of a sample of 466 items from the United States Medical Licensing Examination (USMLE). The present study was conducted in relation to this shared task. For a review of the complete set of submissions to the shared task, please see Yaneva et al. (2024).

2 Related Work

In the last decade, educational assessment has witnessed a surge of interest in predicting item difficulty. Having reviewed 38 papers on item difficulty prediction, AlKhuzaey et al. (2021) provided a summary of the most frequent prediction models used, most studied domains and item types, and features with highest prediction power. ML algorithms such as neural networks and support vector machines (SVM) are commonly employed along with a variety of natural language processing (NLP) techniques used for feature extraction from text data. Language assessment was found to be the most investigated domain, and multiple-choice items were most frequently studied. A greater contribution of AlKhuzaey et al. (2021) lies in its review of the most influential features in item difficulty prediction. While most features can be categorized as either syntactic or semantic, a few studies have used psycholinguistic features (e.g., Pandarova et al., 2019), taking into account the processing of linguistic elements in the brain. The Age of Acquisition (AOA), as one of such "cognitively-motivated" features, offers an index of lexical difficulty based on how early/late in life certain words are acquired (Ha et al., 2019, p. 15). Word concreteness is

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 551–560
June 20, 2024

another psycholinguistic feature. Concrete words are assumed to be processed faster in the brain and thus would expectedly be easier than abstract words (Brysbaert et al., 2014). The use of psycholinguistic features is not a novel approach, however. AOA and word concreteness, among similar features such as word imaginability, have long been on the list of the indices calculated by Coh-Metrix (Bruss et al., 2004). A more recent trend is the use of semantic similarity as a feature, which is discussed further in the following.

Most recently, Štěpánek et al. (2023) compared the performance of several ML algorithms in predicting the item difficulty of reading comprehension tests using features extracted from item texts. Their extracted features include word counts, word frequencies, readability indices, and lexical similarity. For lexical similarity, using Euclidean distance and cosine similarity, they calculated the textual similarity between the question and the correct option as well as between the correct option and the distractors. It was assumed that a higher similarity in the former comparison can make the question easier, while a larger similarity in the latter is associated with higher difficulty (Alsubait et al., 2014). Their results indicated that regularization-based models in general, and the elastic net (RMSE = 0.666) in particular, outperformed other models.

Although we have recently seen an increasing number of attempts to predict item characteristics such as difficulty, the wide range of test domains and other differing contextual factors make it rather difficult to make generalizations across contexts. Therefore, more studies are still needed before more valid conclusions can be drawn regarding the predictability of item characteristics. The purpose of the present study was to contribute to the line of research on predicting item characteristics in medical tests (see, for example, Xue et al., 2020, and Yaneva et al., 2021) by exploring how an assortment of linguistic and non-linguistics features can be utilized along with word embeddings to predict the item difficulty and response time of multiple-choice medical test items.

3 Methods

3.1 Corpus

The corpus of the study is a retired sample of 667 multiple-choice questions from the USMLE. The USMLE is developed by the National Board of Medical Examiners (NBME) and the Federation of State Medical Boards (FSMB) and is administered to both US and Canadian medical students. It consists of three steps, which altogether take nine hours to write. The items are written by experienced medical experts following a set of standardization guidelines. The guidelines help produce high-quality items, the difficulty of which is dependent on the difficulty of the medical content rather than any other extraneous factors.

3.2 Features

A variety of features were extracted mostly from the item stems to be used in our prediction models.

1. **Item Type:** The items in our sample of medical tests can be divided into two groups: text-only and text-and-picture items. Of the 466 items used in the train set, 10.7% (50 items) had a picture supporting the stem text. The use of pictures might help with better and faster understanding of the question.

2. **Exam Part (Step):** As mentioned in the Corpus section, the USMLE has three parts or steps. On average, Step 3 and Step 1 have the highest and lowest item difficulty, respectively. The difference between the exam steps is less considerable in terms of item response time.

3. **Stem Length:** Stem length was measured by counting the number of words in each stem. Longer stems usually take more time to read and understand, and thus they can be more difficult. The stem length of the train set items ranged from 32 to 301 words.

4. **Sentence Length Average:** A very long text can be easy to read if it contains short sentences, while a fairly short text with long sentences can be cumbersome. Therefore, we measured sentence length (as the number of words in a sentence) along with stem length.

5. **Sentence Length Maximum:** Sometimes one single lengthy (or complex) sentence can considerably interfere with comprehension, so we included Sentence Length Maximum as a separate feature in addition to Sentence Length Average.

6. **Option Count:** The higher number of answer options means a higher number of distractors, which is expected to make an item more difficult and time-consuming. Compared to the

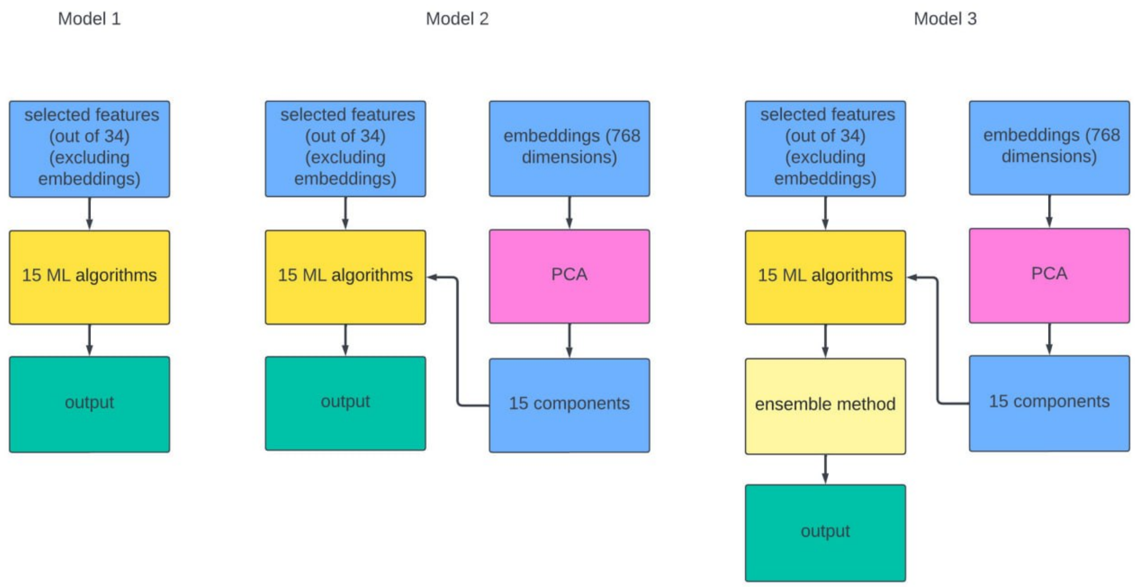

Figure 1: Models 1-3 used for predicting difficulty and response time

minimum of four options, some items in the train set have as many as 10 options. The most common number of options is five.

7. **Challenging Topics:** Based on our observations of several highly difficult items, we formed a short list of potentially more challenging topics in our sample. The list includes the following keywords: ‘kidney’, ‘bleeding’, ‘abdominal’, ‘emergency’, ‘fever’, ‘lung’, ‘abnormality’, and ‘history’. We counted these keywords in lemmatized stems and then assigned each stem a count number accordingly. Items with higher count numbers were expected to be more difficult.

8. **Rare Words Sum:** Less frequent words are usually more difficult (Brysbaert et al., 2011). To calculate the rareness (or difficulty) of the vocabulary of item stems, we looked up each word in the BNC/COCA list (version 2.0.0), a frequency-based list of 25k English words (Nation, 2016). The BNC/COCA list classifies 25k words into 25 frequency groups based on their appearance in the two well-known corpora of BNC (British National Corpus) and COCA (Corpus of Contemporary American English).

9. **Medical Terms Sum:** We used a publicly available list of medical terms (under GNU General Public License v3.0), consisting of terms from two well-known medical dictionaries, namely OpenMedSpel by e-MedTools and Raj&Co-Med-Spel-Chek by Rajasekharan N. of Raj&Co. We counted the number of medical terms in each stem and used that as an indicator of difficulty, assuming that stems with a higher number of medical terms are more difficult and time-consuming to process. Nevertheless, it should be noted that terms can be a double-edged sword, as they can both facilitate the accessibility of information (Baleghizadeh and Yousefpoori-Naeim, 2013) and create obstacles in comprehension (Yousefpoori-Naeim et al., 2018). Moreover, not all terms in a specific domain are equal; they can be placed in a wide range of difficulty (Yousefpoori-Naeim and Baleghizadeh, 2018).

10. **Coh-Metrix Features:** Coh-Metrix is a computational tool that provides 108 indices for text analysis. These indices represent text in terms of its coherence (McNamara et al., 2014). The Coh-Metrix indices used in this study include CNCCaus, CNCTemp, CRFANPa, CRFAO1, CRFCWO1, DESWLlt, LDTTRc, LDVOCDa, LSAGN, LSASSpd, PCCNCz, PCCONNz, PCDCz, PCREFz, PCSYNz, PCTEMPz, RDFRE, SMCAUSlsa, WRDADJ, WRDADV, WRDFRQa, WRDMEAc, WRDNOUN, and WRDVERB. The complete names of these features are provided in Table 2 in the appendix. For more informa-

tion on what each of these features refers to and how they are calculated, see Coh-Metrix version 3.0 indices.

11. **Embeddings:** We used the MPNet encoder to obtain embeddings for each stem text. MPNet is a pre-trained transformer-based language model, which has been shown to outperform similar well-known pre-trained models, such as BERT and RoBERTa, in several tasks (Song et al., 2020). MPNet encoder generates embeddings in the form of 768-dimension vectors. The embeddings represent text in various aspects, including its context, meaning, and syntactic structure.

3.3 Algorithms

We deployed 15 ML algorithms to achieve the highest performance: Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, Stochastic Gradient Descent (SGD), Support Vector Regression (SVR), Decision Tree, Random Forest, Gradient Boosting, Extra Trees, AdaBoost, K-Neighbors, Multilayer Perceptron (MLP), XGBoost, and CatBoost. These algorithms cover a broad spectrum of ML techniques, each with its own strengths and use cases.

3.4 Procedures

Irrespective of the algorithm used, three models were built for each prediction task incrementally. First, a selection of features excluding embeddings was used to train Model 1. Next, embeddings were added to build Model 2. Finally, an ensemble method was utilized to find the best combination of algorithms to be used in Model 3. Figure 1 depicts the structure of the three models in more detail.

Given the high number of our features, we made attempts at different stages of the models to filter out the less relevant features, as feature reduction can enhance model efficiency and lower the risk of overfitting (Ying, 2019). Initially, using a heat map, we detected instances of high correlation in every possible pair of features to address multicollinearity. We marked a correlation coefficient of 0.8 and higher as the presence of multicollinearity (Hae, 2019) and removed one of the two features in the pair. The choice of features for removal was based on theoretical justification and/or literature insights. In a later stage, after Model 1 was initially trained, we gradually removed a few more features based on feature importance results and retrained the model with the truncated list of features. If model performance remained relatively stable, we kept the removed features out of the feature set; otherwise, we re-inserted the removed features one by one to reach comparable performance results. The final lists of selected features used for each prediction task are provided in Table 3 in the appendix. The feature of embeddings went through a reduction process as well. Principal component analysis (PCA) was used to reduce the 768 dimensions of embeddings to 15 components. This number of components was chosen after experimenting with a range of components from 5 to 20, with 15 components yielding the best result.

Cross-validation (CV) was utilized to make the best of the limited data. After randomly leaving 20% of the data out for testing the final models, we ran a 5-fold cross-validation on the remaining 80% subset. Root mean square error (RMSE) results were reported on both the test set and the five folds of the CV subset. A comparison of model performance in training and test sets helps with the detection of overfitting (Ying, 2019).

RMSE was used as the main evaluation metric in the study. It is calculated based on the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{1}$$

where y_i is the actual outcome value and $\hat{y}_i$ is the predicted one for the i-th data, with N denoting the total number of data. RMSE is thus an indicator of the prediction error, i.e., the difference between predicted and actual outcome values. Lower RMSE values indicate lower prediction error.

4 Results

Table 1 presents the RMSE results of all three models in the test and CV subsets for both prediction tasks. In both tasks, Model 2 has a marginally better performance (i.e., lower RMSE) than Model 1, indicating that the addition of embeddings only slightly enhances model performance. Additionally, using the ensemble method (Model 3) did not lead to any performance improvement in either of the tasks.

Unlike the RMSE results, the feature importance results were relatively different in the two prediction tasks. In particular, Coh-Metrix features had a stronger presence in the top features for the difficulty task. In predicting difficulty, PCTEMPz,

Table 1: Model comparisons for predicting difficulty and response time

Models	Difficulty			Response Time		
	Method	**Test RMSE**	**CV RMSE**	**Method**	**Test RMSE**	**CV RMSE**
Model 1	CatBoost	0.277	0.314	K-Neighbors	23.743	25.586
Model 2	AdaBoost	0.269	0.315	K-Neighbors	23.271	24.898
Model 3	Ensemble*	0.269	0.315	Ensemble**	23.271	24.898

*{AdaBoost} **{K-Neighbors}

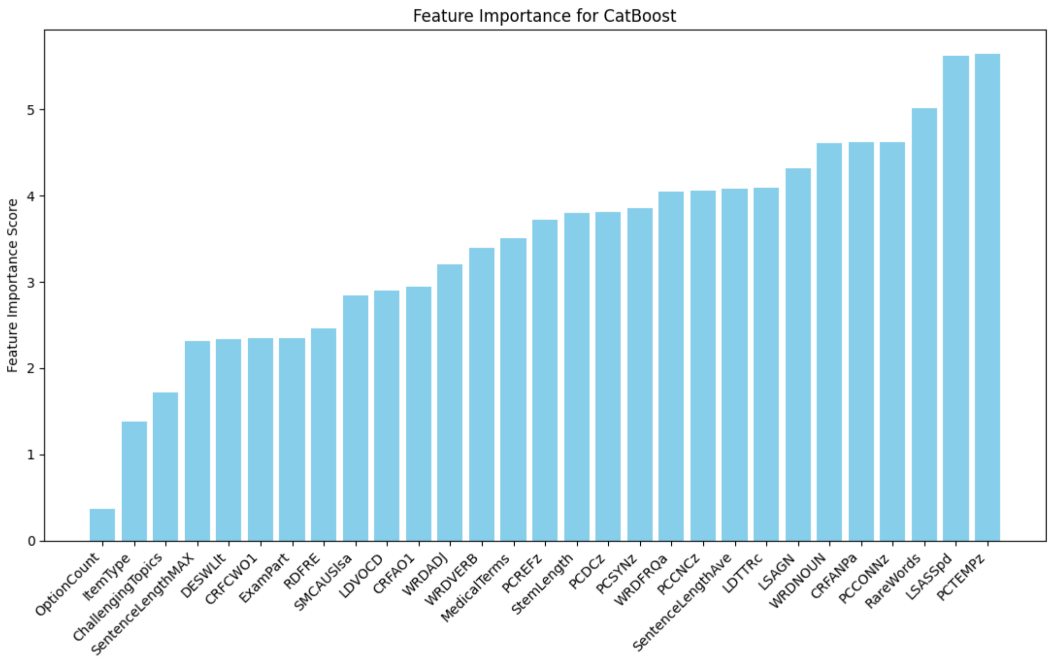

Figure 2: Feature importance scores for predicting difficulty using the CatBoost method

LSASSpd, and Rare Words Sum are the top three features (Figure 2), while Sentence Length Max, Stem Length, and Medical Terms Sum stand out as the top three features predicting response time (Figure 3). Moreover, unlike the prediction task of difficulty, a few Coh-Metrix features were found to have a weak negative relationship with response time.

5 Discussion

A comparison of the RMSE results across the three models in both prediction tasks indicates that the addition of embeddings (i.e., Model 2) had a very small contribution to model performance. While this finding was against our initial expectation, it does bear credence when taking into account the large number of features already fed into Model 1. The selected Coh-Metrix indices coupled with our extracted features (such as Rare Words Sum and Medical Terms Sum) captured most of the variance, leaving not much else to be explained by embeddings. A similar scenario has been present in some other studies. In (Ha et al., 2019), for example, adding either Word2vec or ELMo embeddings to a list of various linguistic features improved RMSE results by minimal margins.

As for Model 3 in both prediction tasks, the ensemble method was ineffective in reducing RMSE values because there was no possible combination of algorithms that would result in a better model performance. In both tasks, the difference between the top-performing algorithm and the rest of the algorithms was wide; therefore, combining the top algorithm with any other one would only harm the performance. Another reason could be that the algorithms are making similar predictions, meaning

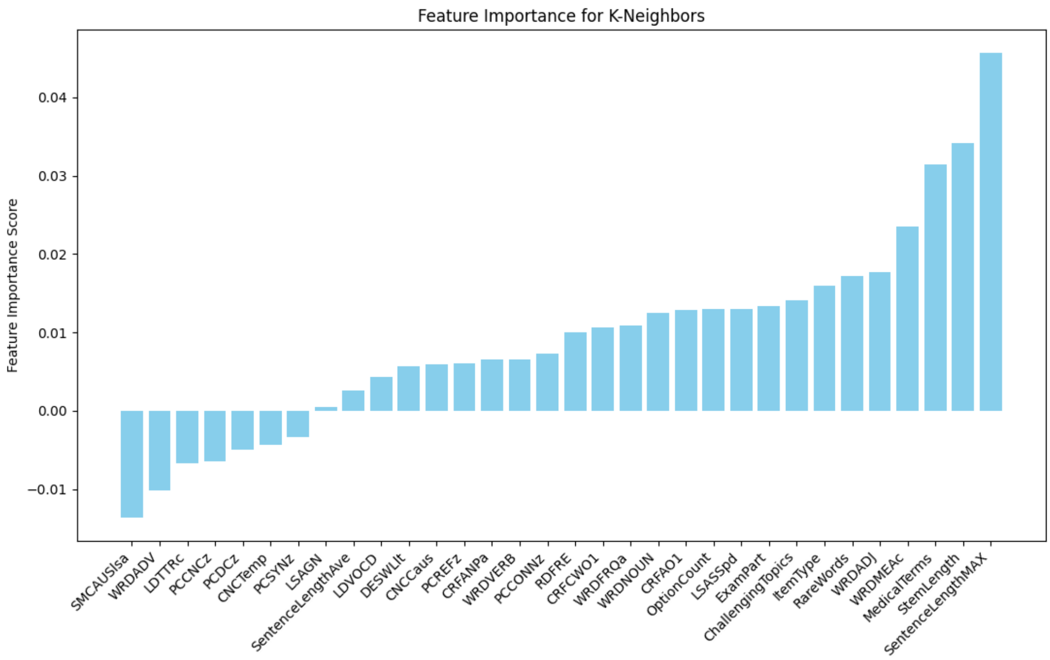

Figure 3: Feature importance scores for predicting response time using the K-Neighbors method

that there is a high correlation between their predictions. The ensemble method usually works best when models trained by different algorithms have different strengths and weaknesses, so combining models could lead to one model compensating for deficiencies in another.

The feature importance scores exhibited dissimilar patterns in the two prediction tasks. Features measuring the cohesion of the stem text played a major role in predicting difficulty: The vast majority of the top predictors of difficulty are cohesion-based Coh-Metrix features. On the other hand, non-Coh-Metrix features, especially length-related ones, constituted the main group of predictors of item response. Length, measured as either the maximum number of words in a sentence (i.e., Sentence Length Max) or the total number of words in the stem text (i.e., Stem Length), is the predominant predictor of item response. Compared to difficulty, response time can be considered less complicated to explain, as it is highly dependent on simple length-related features.

6 Limitations

Two limitations need to be taken into account when interpreting the results of the study. Firstly, the quality of extracted features was dependent on the quality of stem text preprocessing. While preprocessing text data (e.g., tokenization and lemmatization) is generally challenging, the text of medical items can pose additional challenges. The stem of many medical items typically contains a tabulation of data, e.g., laboratory results and blood pressure measures. When embedded within the text, such data can negatively impact the accuracy of feature calculations. For example, a list of items and numbers within a syntactically simple sentence can make it appear as a complex sentence in measures of sentence complexity. It can also interfere with coherence measures calculated through the Coh-Matrix.

The second limitation concerns the results of feature importance. Different algorithms may produce relatively different feature importance sets as they try to reach their highest prediction accuracy. Therefore, the top three or five features in one algorithm can differ from those in another algorithm even with a very close RMSE. To better understand the contribution of each feature to the prediction model, experimental studies are recommended, as the direct effect of individual variables can be more reliably examined through experimental control and manipulation (Yousefpoori-Naeim et al., 2023).

7 Conclusion

The present study explored a selection of diverse features to predict the difficulty and response time of a sample of multiple-choice medical test items using a variety of ML algorithms. In either of the prediction tasks, the addition of embeddings to the list of features did not make a considerable contribution to model performance, and the use of the ensemble method was not effective either. In feature importance scores, however, the two tasks showed dissimilar patterns. Features measuring cohesion were especially effective in predicting difficulty, while length-related features were the main predictors of response time.

While future studies can examine the role of many other features in predicting item characteristics of medical tests, we would like to draw attention to collecting data from item writers to be used as a potential feature. Especially in the case of item difficulty, medical test writers can be asked to rate the difficulty of the items they develop. While students might perceive items differently from what test writers would assume, item writers' ratings could still correlate highly with actual difficulty values. This feature enjoys high practicality and low cost, as item writers can give difficulty ratings as they write their own items. A more advanced, but also more expensive approach is to have item writers rate each others' items as well.

References

Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2021. A systematic review of data-driven approaches to item difficulty prediction. In *Artificial Intelligence in Education*, pages 29–41, Cham. Springer International Publishing.

Tahani Alsubait, Bijan Parsia, and Uli Sattler. 2014. Generating multiple choice questions from ontologies: Lessons learnt. In *OWLED*, pages 73–84. Citeseer.

Margit Antal. 2013. On the use of elo rating for adaptive assessment. *Studia Universitatis Babes-Bolyai, Informatica*, 58(1):29–41.

Sasan Baleghizadeh and Mehrdad Yousefpoori-Naeim. 2013. Surveying metalanguage through three efl textbooks. *E-International Journal of Educational Research*, 4(3):27–40.

Michell Bruss, Michael J. Albers, and Danielle McNamera. 2004. Changes in scientific articles over two hundred years: a coh-metrix analysis. In *Proceedings of the 22nd Annual International Conference on Design of Communication: The Engineering of Quality Documentation*, SIGDOC '04, page 104–109, New York, NY, USA. Association for Computing Machinery.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental Psychology*, 58(5):412–424. PMID: 21768069.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Charles DePascale and Brian Gong. 2020. Comparability of individual students' scores on the "same test. *Comparability of large-scale educational assessments: Issues and recommendations*, pages 25–48.

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, Florence, Italy. Association for Computational Linguistics.

Kim Jong Hae. 2019. Multicollinearity and misleading statistical results. *kja*, 72(6):558–569.

Patrícia Martinková and Adéla Hladká. 2023. *Computational aspects of psychometric methods: With R*. CRC Press.

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.

Ian Stephen Paul Nation. 2016. *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Irina Pandarova, Torben Schmidt, Johannes Hartig, Ahcène Boubekki, Roger Dale Jones, and Ulf Brefeld. 2019. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, 29:342–367.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, Seattle, WA, USA → Online. Association for Computational Linguistics.

Victoria Yaneva, Daniel Jurich, Le An Ha, and Peter Baldwin. 2021. Using linguistic features to predict the response process complexity associated with answering clinical MCQs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building*

Educational Applications, pages 223–232, Online. Association for Computational Linguistics.

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.

Xue Ying. 2019. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022.

Mehrdad Yousefpoori-Naeim and Sasan Baleghizadeh. 2018. Towards finding a difficulty index for english grammatical terminology. *Terminology*, 24(2):236–261.

Mehrdad Yousefpoori-Naeim, Okan Bulut, and Bin Tan. 2023. Predicting reading comprehension performance based on student characteristics and item properties. *Studies in Educational Evaluation*, 79:101309.

Mehrdad Yousefpoori-Naeim, Lawrence Jun Zhang, and Sasan Baleghizadeh. 2018. Resolving the terminological mishmash in teaching link words in efl writing. *Chinese Journal of Applied Linguistics*, 41(3):321–337.

Lubomír Štěpánek, Jana Dlouhá, and Patrícia Martinková. 2023. Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19).

A Appendix

Table 2: Coh-Metrix feature labels and their descriptions

Feature Label	Description
CNCCaus	Causal connectives incidence
CNCTemp	Temporal connectives incidence
CRFANPa	Anaphor overlap, all sentences
CRFAO1	Argument overlap, adjacent sentences binary, mean
CRFCWO1	Content word overlap, adjacent sentences proportional, mean
DESWLlt	Word length, number of letters, mean
LDTTRc	Lexical diversity, type-token ratio, content word lemmas
LDVOCDa	Lexical diversity, VOCD, all words
LSAGN	LSA given/new, sentences, mean
LSASSpd	LSA overlap, all sentences in a paragraph, standard deviation
PCCNCz	Text Easability, PC Word concreteness, z score
PCCONNz	Text Easability, PC Connectivity, z score
PCDCz	Text Easability, PC Deep cohesion, z score
PCREFz	Text Easability, PC Referential cohesion, z score
PCSYNz	Text Easability, PC Syntactic simplicity, z score
PCTEMPz	Text Easability, PC Temporality, z score
RDFRE	LSA verb overlap
SMCAUSlsa	Flesch Reading Ease
WRDADJ	Adjective incidence
WRDADV	Adverb incidence
WRDFRQa	CELEX Log frequency for all words, mean
WRDMEAc	Meaningfulness, Colorado norms, content words, mean
WRDNOUN	Noun incidence
WRDVERB	Verb incidence

Table 3: List of features used in each prediction task

Features	Difficulty	Item Response
Item Type	•	•
Exam Part (Stem)	•	•
Stem Length	•	•
Sentence Length Average	•	•
Sentence Length Maximum	•	•
Option Count	•	•
Challenging Topics	•	•
Rare Words Sum	•	•
Medical Terms Sum	•	•
CNCCaus		•
CNCTemp		•
CRFANPa	•	•
CRFAO1	•	•
CRFCWO1	•	•
DESWLlt	•	•
LDTTRc	•	•
LDVOCDa	•	•
LSAGN	•	•
LSASSpd	•	•
PCCNCz	•	•
PCCONNz	•	•
PCDCz	•	•
PCREFz	•	•
PCSYNz	•	•
PCTEMPz	•	
RDFRE	•	•
SMCAUSlsa	•	•
WRDADJ	•	•
WRDADV		•
WRDFRQa	•	•
WRDMEAc		•
WRDNOUN	•	•
WRDVERB	•	•