Medical Item Difficulty Prediction Using Machine Learning

Hope Adegoke[1], **Ying Du**[2], **Andrew Dwyer**[2]
[1]Educational Research Methodology, University of North Carolina, Greensboro.
[2]American Board of Pediatrics, 111 Silver Cedar Court, Chapel Hill, NC 27514.
[1]hoadegoke@uncg.edu, [2]ydu@abpeds.org, [2]adwyer@abpeds.org

Abstract

This study examines the prediction of medical exam item difficulty using NLP and machine learning. A dataset of 1,154 MOCA-Peds items was represented with lexical features, cohesion metrics (TAACO), blueprint encodings, BioMedBERT embeddings, semantic similarity, and unsupervised semantic indicators. Regression models predicted IRTb parameters with RMSE $\approx$ 1.37 and $R^2 \approx .29$, a $\sim 15\%$ improvement over baseline. Ordinal classifiers reached 30–34% accuracy, doubling the baseline (17%), with $\sim$68% adjacent accuracy. Feature importance showed blueprint, item structural features (number of options, option length), and semantic embeddings as the strongest predictors. These findings demonstrate the feasibility of difficulty prediction in medical assessments and highlight the value of combining domain-specific semantics with latent difficulty signals for efficient test development.

1 Introduction

Having an idea of the difficulty of items before operational use supports efficient test construction, helps to achieve target score precision, and reduces the cost and security risks of pre-testing (Settles et al., 2020). Previous work using text features alone has shown mixed results, often only modestly surpassing naive baselines (Štepánek et al., 2023). Shared-task evidence suggests that transformers can help when carefully tuned and combined with complementary features (Yaneva et al., 2024; Gombert et al., 2024). In medical licensing contexts where items require specialized knowledge and clinical reasoning, prediction of difficulty has been particularly challenging (Ha et al., 2019). We extend this literature by integrating different categories of linguistic and semantic feature families, including domain-specific biomedical embeddings (Gu et al., 2021), and by encoding content blueprints, then evaluating both continuous and ordinal formulations of difficulty.

2 Related Works

Early studies of automated item difficulty prediction focused on surface-level text features such as length, readability, and lexical counts, but these typically explained little variance (McNamara et al., 2014; Štepánek et al., 2023; Ha et al., 2019). With the rise of NLP, transformer-based models achieved top performance in the BEA 2024 Shared Task on difficulty prediction (Gombert et al., 2024), while methods using traditional linguistic/cohesion features (including clinical embeddings and principal component features) also showed competitive performance though generally lower than the top transformer-based approaches (Yaneva et al., 2024; Tack et al., 2024).

Medical assessments present unique challenges. Ha et al., 2019 reported only modest gains over baseline for medical MCQs, reiterating the importance of incorporating domain expertise. Domain-specific embeddings like BioMedBERT (Gu et al., 2021) provide richer representations of medical terminology. Furthermore, content-based encodings such as blueprints or cognitive-level taxonomies (Anderson and Krathwohl, 2001) can serve as structured priors for difficulty. Our work extends these lines by combining linguistic, semantic, and content features, evaluating both continuous prediction of IRTb and ordinal classification into difficulty bins.

3 Method

3.1 Data & Targets

We analyze 1,154 multiple-choice items from the Maintenance of Certification Assessment for Pediatrics (MOCA-Peds), a longitudinal, online assessment that allows pediatricians to demonstrate ongoing knowledge through periodic testing rather

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 185–190

than a single high-stakes exam (Leslie et al., 2018). Each item consists of a clinical stem and four or five answer options (A–D or A–E). Every item has a calibrated IRT difficulty parameter (b), which serves as the outcome variable for the prediction. In the regression setting, the task is to predict the continuous IRTb value directly. In the classification setting, items are grouped into five ordinal difficulty categories by dividing the IRTb distribution into quintiles, ranging from the easiest to the hardest items. Because there are five equally sized bins, random guessing of the algorithms would yield a uniform-chance accuracy of 20% for predicting the exact difficulty class.

3.2 Features

We construct five feature families:

Lexical features: Stem and option lengths (in both characters and tokens), punctuation counts, total number of options (4 and 5), and lexical overlap indices. Overlap was measured with Jaccard similarity between the stem and the correct option, and between the stem and the distractors.

TAACO cohesion features: Cohesion and readability indices computed with the Tool for the Automatic Analysis of Cohesion - TAACO (Crossley et al., 2016, 2019). These include lexical diversity, referential overlap, connectives usage, and standard readability metrics.

BioMedBERT semantic features: Mean-pooled contextual embeddings of the stem and options from BioMedBERT (Gu et al., 2021). From these embeddings, we derived cosine similarities (stem–correct option and stem–distractors), dispersion among option embeddings (mean pairwise cosine distance), and principal components of the stem embeddings to provide lower-dimensional semantic factors.

Blueprint encoding: Smoothed target encoding of Level-2 content domains from the MOCA-Peds blueprint. Each item's category was assigned the smoothed mean difficulty from training folds, providing a structured content-based prior.

Unsupervised difficulty features: Embedding-derived features that do not use the target difficulty, such as stem "uniqueness" relative to the corpus, cluster-based indicators (cluster size, distance to centroid), and coherence/variance scores. These are designed to capture latent difficulty signals without relying on calibrated b parameters.

3.3 Modeling & Validation

We approach difficulty prediction through two complementary pipelines: regression for continuous IRTb values and ordinal classification for quantile-based difficulty bins. For both pipelines, we used a nested cross-validation strategy. In the outer loop, five folds of data ensured every item served once as test data to estimate generalization error. Within each training partition, randomized hyperparameter search with inner cross-validation selected the best model configuration. This setup prevents information leakage between tuning and evaluation, which produces a robust and unbiased performance estimates (Varma and Simon, 2006; Arlot and Celisse, 2010). For regression, we combined interpretable linear models with flexible non-linear methods. Elastic Net and Ridge provided regularized linear baselines (Štepánek et al., 2023), while Random Forest and gradient boosting methods (XGBoost, LightGBM, CatBoost) captured non-linear interactions. Evaluation emphasized Root Mean Square Error (RMSE), supplemented by Mean Absolute Error (MAE), R^2, and Spearman rank correlation to reflect both the magnitude and the ranking of difficulty. Calibration slopes were also inspected to assess systematic under- or overestimation.

RMSE (Root Mean Square Error): average magnitude of prediction errors, penalizing larger errors more heavily.

MAE (Mean Absolute Error): average absolute difference between predicted and true values.

R^2 (Coefficient of Determination): proportion of variance in item difficulty explained by the model.

Spearman's ρ: correlation between predicted and true rankings of item difficulty, reflecting how well the model preserves ordering.

Calibration Slope: regression slope of observed versus predicted difficulty; values close to 1 indicate unbiased, well-scaled predictions.

For classification, we tested Logistic Regression, SVMs, and Random Forest, consistent with prior work on medical MCQ difficulty prediction (Ha et al., 2019). We additionally evaluated gradient boosting classifiers (XGBoost, LightGBM, CatBoost), which have been shown to be competitive in recent difficulty prediction tasks (Yousefpoori-Naeim et al., 2024). We reported overall accuracy, macro-F1, Ordinal Mean Absolute Error, Adjacent Accuracy (crediting predictions within one difficulty level), and Top-2 Accuracy. Such metrics are

recommended in ordinal classification/assessment prediction settings, though prior work uses subsets of them (Ha et al., 2019; Gombert et al., 2024).

Accuracy: proportion of items where the predicted bin exactly matched the true bin.

Macro-F1: unweighted average of F1-scores across bins, combining precision and recall.

Ordinal MAE: mean absolute difference between predicted and true bin indices, capturing distance on the ordered scale.

Adjacent Accuracy: percentage of predictions that were exact or within one difficulty level.

Top-2 Accuracy: percentage of items where the correct bin was among the model's two highest-scored predictions.

All models were implemented in standardized Python pipelines with preprocessing, scaling, and leakage-safe encoding. Embeddings were precomputed and cached to ensure efficiency, and fixed seeds were used for reproducibility.

4 Results

4.1 Regression (predicting IRTb)

In the regression task, models were trained to predict continuous IRT b parameters directly from item features. The baseline predictor, which always returned the mean item difficulty, yielded RMSE $\approx$ 1.63 and explained virtually none of the variance ($R^2 \approx 0$). In contrast, all feature-based models substantially outperformed this baseline. As shown in Table 1, the best results were achieved by Elastic Net and Random Forest, which attained RMSE $\approx$ 1.37 and explained about 28–29% of variance ($R^2 \approx 0.28$–0.29). Both also showed moderate rank-order correlations with true difficulty ($\rho \approx 0.45$), indicating that they not only approximated difficulty values but also captured relative ordering among items. Gradient boosting methods (LightGBM and CatBoost) performed nearly as well, with RMSE $\approx$ 1.37 and $R^2 \approx 0.288$. Ridge regression was slightly weaker ($R^2 \approx 0.276$), while XGBoost lagged behind with the highest RMSE ($\approx$ 1.40) and the lowest explained variance ($R^2 \approx$ 0.26).

Calibration analysis confirmed that model predictions were well aligned with observed values: Random Forest achieved a slope close to 1.0, while Elastic Net slightly underestimated extreme difficulties (slope $\approx$ 1.1). In practical terms, given the observed range of b values ($\approx$ –5 to +5), the error reduction from 1.63 to 1.37 translates into roughly a 15% gain in predictive precision from the baseline.

4.2 Ordinal classification (5 difficulty bins)

For the 5-class ordinal classification task, models substantially outperformed the baselines. The majority-class baseline reached only 17.2% accuracy, while a uniform random predictor would achieve $\approx$20% accuracy by chance.

As shown in Table 2, the Random Forest classifier achieved the strongest performance with 34.5% accuracy, approximately double the majority-class baseline. Its Macro-F1 (0.35) was aligned with accuracy, reflecting fairly uniform performance across difficulty bins. The ordinal-specific metrics confirmed its usefulness: the Ordinal Mean Absolute Error was 1.14 (vs. 1.48 for baseline), and Adjacent Accuracy reached 68%, indicating that two-thirds of predictions were either exact or within one difficulty level. The Top-2 accuracy of 54% further shows that the true class was frequently among the two highest-scored bins.

LightGBM and XGBoost followed closely (overall accuracy $\approx$0.34 and $\approx$0.34 respectively), while SVM (RBF kernel) and Logistic Regression trailed modestly ($\approx$0.30–0.32 overall accuracy). Importantly, even the weaker models still exceeded baseline performance, confirming that item features contain reliable ordinal difficulty signals.

4.3 Feature Importance

We examined feature importance across regression and classification models. Figure 1 show us that for Random Forest regression, the strongest predictors were the blueprint encoding, the number of answer options, and the length of option E. Each accounted for around 10% of the model's explanatory variance, confirming that both content area and item format influence difficulty.

Embedding-based features also played a key role. Several principal components from BioMedBERT stem embeddings and stem–distractor cosine similarities ranked among the top predictors, indicating that semantic complexity and distractor plausibility strongly shaped difficulty. In contrast, cohesion indices from TAACO and traditional readability measures contributed little when richer semantic and content features were available.

For the ordinal classification Random Forest, the same pattern emerged: option E length and blueprint encoding dominated, followed by embedding-derived factors and unsupervised simi-

Table 1: Cross-validated regression results (5-fold outer CV). Lower RMSE is better.

Model	RMSE	MAE	R^2	Spearman ρ	Calib. slope
Elastic Net	**1.3719**	1.0759	**.2881**	0.4584	1.10
Random Forest	**1.3685**	1.0744	**.2916**	0.4464	0.98
LightGBM	1.3715	1.0763	.2885	0.4577	1.12
CatBoost	1.3717	1.0752	.2882	0.4530	1.07
Ridge	1.3838	1.0852	.2757	0.4400	0.96
XGBoost	1.3976	1.1034	.2612	0.4174	1.04
Baseline (mean)	1.6262	1.2685	∼0.00	∼0.00	—

Table 2: Ordinal classification results (5-fold outer CV).

Model	Accuracy	Macro-F1	Ordinal-MAE	Adjacent Acc.	Top-2 Acc.
Random Forest	**.345**	**.347**	**1.136**	**.679**	.538
LightGBM	.343	.338	1.187	.657	**.555**
XGBoost	.337	.335	1.156	.677	.536
SVM (RBF)	.322	.324	1.174	.666	.538
Logistic Reg.	.309	.296	1.247	.638	.524
Baseline (maj.)	.172	.150	1.478	.546	.382

larity features (e.g., stem–nearest neighbor cosine distance). This shows that difficulty is not just a function of surface text complexity but is rooted in the interaction of content domain, item structure, and semantic relationships among options.

5 Discussion

Our models achieved modest but meaningful predictive power: regression explained about 28–29% of variance in IRTb, and ordinal classification reached 34% accuracy with 68% adjacent accuracy. These gains over baseline suggest that automated difficulty prediction can support item development, though the strength of the predictions remains limited.

Importantly, our results are consistent with prior studies. Štepánek et al., 2023 reported similar variance explained when predicting reading comprehension item difficulty, and Ha et al., 2019 found only modest gains for medical MCQs. Recent BEA-2024 shared-task findings (Yaneva et al., 2024; Gombert et al., 2024; Tack et al., 2024; Yousefpoori-Naeim et al., 2024) likewise show that even transformer-based systems reach only moderate correlations, underscoring a common ceiling in this line of work.

The drivers of prediction in our study, which are option structure, blueprint encoding, and biomedical embeddings, mirror some of those highlighted in other research (Ha et al., 2019; Tack et al., 2024). Readability and cohesion features offer little contribution once richer, domain-specific features are available. This points to why the ceiling persists: difficulty depends not just on text but also on broader context, reasoning steps, and examinee knowledge, factors not fully captured by textual features.

From a practical perspective, these models are best used for screening and triage: flagging potentially too-easy or too-hard items, or giving item writers feedback about content areas and option structures. They are unlikely to replace psychometric calibration (at least not yet), but can reduce workload and guide review.

Looking ahead, progress will likely come from incorporating richer modalities (stimuli, visuals), domain-adapted embeddings, and design-aware features that better align with the cognitive processes behind item difficulty. Until then, automated prediction should be seen as an assistive tool that complements, rather than substitutes the current process.

6 Conclusion

This study shows that predicting the difficulty of medical multiple-choice items is feasible when models combine diverse linguistic, semantic, and content-informed features. By integrating domain-specific biomedical embeddings and blueprint encodings alongside lexical and cohesion measures, our models achieved measurable improvements over baselines in both continuous and ordinal formulations of difficulty. Importantly, the results highlight that difficulty prediction is not driven by surface text length alone but by deeper signals of what the item is about and how it is structured.

The practical implication is that automated prediction can serve as a support tool in item devel-

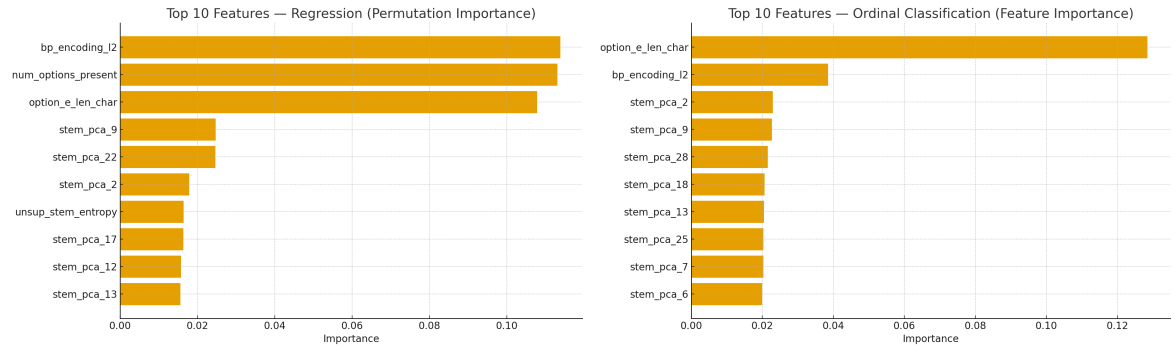

Figure 1: Top 10 predictive features identified by importance analysis. Left: Regression models predicting continuous IRTb. Right: Ordinal classification models predicting difficulty bins.

opment pipelines, sorting items for expert review, guiding test assembly, and reducing reliance on costly pretesting. At the same time, the modest explanatory power and the prominence of dataset-specific signals (such as Option E length) remind us that these models should complement, not replace, expert judgment and psychometric validation.

Future work should extend these methods to larger and more diverse item pools, explore fine-tuned transformer models, and incorporate additional sources of difficulty such as cognitive complexity ratings (Bloom's taxonomy) or multimedia elements. Taken together, the findings provide evidence that machine learning can play a constructive role in modern test development, enhancing efficiency while respecting the central role of human expertise.

7 Limitations and future work

This study was limited by the use of a single dataset of 1,154 MOCA-Peds items, which may constrain generalizability. Some highly ranked predictors, such as Option E length, applied to only a small fraction of items (88/1154) and may reflect dataset-specific patterns rather than universal drivers of difficulty. In addition, the models considered only text and blueprint features, without incorporating multimedia content or group-level differences.

Future work should validate these findings across larger and more diverse item banks, explore explicit cognitive-level annotations, and investigate fine-tuned transformer models trained on exam text. Embedding difficulty prediction into item development workflows to provide real-time feedback to item writers is a promising application.

References

Lorin W. Anderson and David R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. Taaco 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Simon Gombert, Lasse Menzel, Dimitrios Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 483–492.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Jenny 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.

Laurel K Leslie, Murrey G Olmsted, Adam Turner, Carol Carraccio, Andrew Dwyer, and Linda Althouse. 2018. Moca-peds: development of a new assessment of medical knowledge for continuing certification. *Pediatrics*, 142(6):e20181428.

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.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine-learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. ITEC at BEA 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521, Mexico City, Mexico. Association for Computational Linguistics.

Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91.

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

Mohammad Yousefpoori-Naeim, Shirin Zargari, and Zahra Hatami. 2024. Using machine learning to predict item difficulty and response time in medical tests. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 551–560, Mexico City, Mexico. Association for Computational Linguistics.

Lukáš Štepánek, Jana Dlouhá, and Patrícia Martinková. 2023. Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, 11(19):4104.