Review of Text-Based Approaches to Item Difficulty Modeling in Large-Scale Assessments

Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, Tianyi Zhou
University of Maryland
sjpeters@umd.edu, hjiao@umd.edu

Abstract

Item difficulty plays a crucial role in evaluating item quality, test form assembly, and interpretation of scores in large-scale assessments. Traditional approaches to estimate item difficulty rely on item response data collected in field testing, which can be time-consuming and costly. To overcome these challenges, text-based approaches leveraging machine learning and natural language processing, have emerged as promising alternatives. This paper reviews and synthesizes 37 articles on automated item difficulty prediction in large-scale assessments. Each study is synthesized in terms of the dataset, difficulty parameter, subject domain, item type, number of items, training and test data split, input, features, model, evaluation criteria, and model performance outcomes. Overall, text-based models achieved moderate to high predictive performance, highlighting the potential of text-based item difficulty modeling to enhance the current practices of item quality evaluation.

1 Introduction

Large-scale assessments are often used to make high-stakes decisions such as grade promotion, professional certification, or college admission, so they must adhere to professional standards for test development to ensure validity, reliability, and fairness (AERA, APA, & NCME, 2014). The most common approach for estimating item difficulty has conventionally been conducted through field-testing, where newly created items are embedded in an operational test form. These items are used to collect item response data to estimate item parameters (e.g., difficulty) using classical test theory (CTT) or item response theory (IRT) and they are not used for scoring (Benedetto, 2023). Despite its ability to yield accurate item difficulty estimates, this approach has been criticized for being time-consuming and costly (AlKhuzaey et al., 2024; Hsu et al., 2018). Another approach for estimating item difficulty has been through expert ratings, though this is seldom used in developing large-scale assessments due to its subjective nature.

To address these limitations, text-based approaches for item difficulty prediction have offered a fast, objective, and scalable alternative. The timeline of these approaches followed a few noticeable trends. In the early stages, the literature was dominated by feature-based approaches that relied on manually defined variables that are hypothesized to influence item difficulty (e.g., Loukina et al., 2016; Perkins et al., 1995). Later, studies started to include word embeddings, which are numeric vectors representing the semantic relationships among words (e.g., Hsu et al., 2018). With the development of deep learning, embeddings were also extracted from deep neural networks, considering how words and phrases interact within the context of the text (e.g., Xue et al., 2020). Most recently, since 2020, transformer-based models, including small language models (SLMs) and large language models (LLMs), have been utilized, capturing nuanced semantic and contextual relationship (e.g., Li et al., 2025; Tack et al., 2024). These models have the potential to improve model predictive performance, but it comes at the cost of interpretability.

The goal of the present review is to highlight the recent developments in the use of machine learning and language-model based approaches for item difficulty prediction, with a focus on large-scale assessments. Several research questions guide our investigation: (1) What text-based methods, especially advanced language model-based approaches, were applied to predict item difficulty? (2) What domains and item types were most frequently investigated? (3) Which text-based features were most frequently investigated in classic machine learning models? (4) Which

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 37–47

evaluation criteria were used to assess model performance? (5) What was the distribution of evaluation outcomes? What does this reveal about the typical range and variability in item difficulty prediction modeling performance?

2 Related Work

The exploration of text-based approaches to model item difficulty has been ongoing for decades and a few studies have synthesized the research findings in a systematic way. Ferrara et al. (2022) conducted a domain-specific review that summarized 13 item difficulty modeling studies, focusing on high-stakes reading comprehension exams. This review found that statistical models such as ordinary least squares regression were utilized in every study and only two studies employed natural language processing (NLP) techniques. These findings highlight the emerging but still limited text-based methods for item difficulty estimation.

More recent reviews included articles that employed advanced models, which rapidly emerged from the mid-2010s (e.g., AlKhuzaey et al., 2024; Benedetto et al., 2023). Benedetto et al. (2023) conducted a narrative review of the literature and focused on approaches for question difficulty estimation from text from 38 studies published between 2015-2021. They provided a structured taxonomy to organize the approaches and analyzed the most effective methods in different scenarios. Results showed that, in general, simple models leveraging linguistic features performed just as well as end-to-end neural networks for language assessments; but for other subject domains (e.g., math, science) end-to-end neural networks, especially transformers led to increased performance. Their findings also highlighted a shift from readability and word-complexity features-based classic machine learning models to modern deep learning-based, NLP approaches.

AlKhuzaey et al. (2024) conducted a systematic review of 55 item difficulty prediction articles that placed no constraint on time frame, resulting in coverage from the years 1995 to 2022. Compared to previous reviews, they extended the scope to include an in-depth analysis of the most frequently investigated content domains, difficulty parameters, model features, models, input, item types, evaluation metrics, and the number of publications produced over the years. The results highlighted that linguistic play a critical role in estimating item difficulty, syntactic features are frequently captured using NLP tools to count textual elements, and with the development of neural language models, semantic features were increasingly explored.

Similarly, Luecht (2025) summarized years of item difficulty modeling research through 2022. The author explains that item difficulty modeling has evolved along two pathways: the strong theory pathway and the statistical control pathway. The strong theory pathway was most prevalent in the early years of item difficulty modeling research, and it is based on the idea that item design choices should be grounded in strong cognitive and learning theories. With the rise of machine learning and NLP-based text analytics, there has been a gradual shift to the statistical control pathway, that aims to identify variables that empirically explain item difficulty. Under this framework, the primary focus is improving model prediction performance, rather than aligning with cognitive theory.

Though AlKuzaey et al. (2024) and Bendetto et al. (2023) provided an in-depth summary of automated item difficulty prediction methods, they share similar limitations. All included articles were published no later than 2022 and they did not focus on large-scale assessments. Additionally, AlKhuzaey et al. (2024) included articles that used expert ratings as ground truth difficulty, but this is not a valid approach for item difficulty estimation in large-scale assessments due to subjectivity and inconsistency.

Given that language model-based approaches have vastly developed in the past three years (2023-2025), an updated synthesis of the literature is warranted. Another unique contribution of our review is the reporting of model performance outcomes, including the distribution of values obtained across evaluation metrics. This can act as a useful reference for future research by providing reference points for evaluating model performance.

3 Methods

We conducted a comprehensive literature search for articles published through May 2025 across multiple databased, including Google, Google Scholar, IEEE Xplore, ArXiv, Scopus, Springer, and ERIC. Additional searches were performed on the websites of the National Council on Measurement in Education (NCME) and a relevant competition platform (i.e., the NBME Item Difficulty Prediction Competition) to locate papers

submitted by participants. A Boolean search strategy was employed using keyword combinations in full text: *(item OR question) AND difficulty AND (AI prediction OR prediction using machine learning OR automatic prediction OR modeling).*

After an initial screening based on titles, 93 articles were identified. Next, 17 articles were excluded after reviewing the abstract and keywords for relevance, resulting in 76 articles. The full text of these articles was screened and 52 articles were excluded based on one or more of the following reasons: (1) the assessment was not large-scale, (2) the study focused on text complexity or readability rather than item difficulty, (3) the study did not focus on automated prediction based on item text (4) the article was a review, or (5) the item difficulty parameter was not obtained from item responses from human test-takers. A total of 24 articles remained for in-depth analysis.

Later, a forward hand-search was conducted to ensure that all related articles have been comprehensively included. For each included eligible article, we found all subsequent articles that cited it and conducted another round of screening. In this procedure, 19 additional articles were found, and after screening following the same exclusion criteria listed above, 13 articles were included in the review. In total, 37 articles were coded and analyzed for this review, consisting of conference papers (n = 20), journal articles (n = 7), research reports (n = 3), pre-prints (n = 5), and master's or doctoral theses (n = 2).

Since there could be more than one dataset or difficulty parameter analyzed in one paper, we treat these as separate studies. Consequently, 46 studies resulted from the 37 articles. To differentiate the number of articles from the number of studies, we used n for the number of articles and k for the number of studies, hereafter.

For each study we record the article information including title, authors, and publication year; dataset name, difficulty parameter, subject domain, item type, number of items, train and test dataset split, engineered features, models, evaluation criteria, and model performance. Descriptive analyses were performed, and results were reported using count-based aggregation and percentages. Model performance outcomes for each evaluation criterion with sufficient data across studies were summarized using descriptive statistics including minimum, maximum, median, mean, and standard deviation.

4 Results

4.1 Publication Year

Automated item difficulty prediction has come in two waves: one in the mid 1990s, and another beginning in the early 2010s. The resurgence is likely related to the peak of automated question generation research and the rise of computerized adaptive testing around 2014 to 2018 (AlKhuzaey et al., 2024; Kurdi et al., 2021), since item difficulty modeling is essential to evaluate the quality of newly created items. Ever since then research on this topic has been on the rise, with a large spike in 2024 due to the Building Educational Applications (BEA) shared task on automated item difficulty prediction and response time that launched in June 2024.

4.2 Item Difficulty Parameter

When the item difficulty parameter is a continuous parameter, item difficulty prediction is framed as a regression problem. In contrast, when it is defined using categorical levels (e.g., easy, medium, hard), it becomes a classification task (e.g., Hsu et al., 2018). In the context of large-scale exams, it was found that most item difficulty studies predicted a continuous value, which is consistent with the common practice of representing item difficulty in terms of either p-values or IRT b-parameters. Specifically, the most frequently reported methods were IRT b-parameter (k=14, 30.43%), transformed p-value (k=11, 23.91%), and traditional p-value (k=9, 19.57%). Other approaches including categorical difficulty levels (k=5, 10.87%), error rate (k=4, 8.70%), and Delta (k=3, 6.52%), were less common.

4.3 Subject Domain

Test subject domains included language proficiency (k = 23, 50.00%), medicine (k = 15, 32.61%), math (k = 4, 8.70%), science (k = 2, 4.35%), analytical reasoning (k = 1, 2.17%), and social studies (k = 1, 2.17%). Language proficiency and medicine dominate the literature likely due to that the high volume of large-scale exams in these domains made the data publicly available.

4.4 Item Type

Counting each item type once per study, a total of 60 item types were identified across the reviewed materials because several articles examined multiple item types within the same study. Multiple choice (MC) items accounted for the largest share, appearing 38 times (63.33%), followed by fill-in-the-blank reported eight times (13.33%), constructed-response items reported four times (6.67%), and matching items reported twice (3.33%). Each of the following item types were only reported once: complete-the-forms, notes, table, flowchart, or summary, complete-the-table, label the diagram, plan, or map, true/false, classifying, and sorting (3.33% each).

4.5 Number of Items

There was a wide range (348 to 106,210) in the number of items that were used across studies, showing great variability in dataset size. Most studies (k=17) used datasets between 500 and 2,000 items, largely because 11 studies used the data from the BEA shared task with 667 items. Only two studies used a very large dataset with more than 30,000 items (i.e., RACE++ (106,210) used in Benedetto, 2023; IFLYTEK (30,817) used in Huang et al., 2017)[1].

4.6 Training and Test Dataset Split

To develop different models for item difficulty prediction, a dataset is often divided into training, validation, and test datasets. The training set is used to learn patterns and relationships in the data, validation is used to fine-tune the model, and test is used to evaluate model performance on unseen data. Not all studies reported the percentage of data used for validation, so for consistency, training and validation percentages were combined for studies with three splits. We also note that several studies experimented with multiple train and test dataset splits (Benedetto, 2023; Bulut et al., 2024; Huang et al., 2017).

A wide variety of train/test data splits were observed. Reported as percentages they include: 40/60, 50/50, 60/40, 70/30, 75/25, 80/20, 83/17, 84/16, 85/15, 90/10, 93/7, and 95/5. The most common dataset split was 70% for training and 30% for testing, reported 14 times (28.00%), followed by 80/20 reported six times (12.00%), and 50/50, 90/10, and 95/5 reported three times each (6.00%, each). The remaining train and test data split combinations only appeared in two studies or less, and nine studies (18.00%) did not report this information.

4.7 Input

The input used to train the model refers exclusively to the original, unprocessed components of the item (i.e., item stem (lead-in and/or questions), correct answer, distractors, figures or reading passages when applicable). Again, some studies experimented with multiple combinations, and each was counted once per study, for a total count of 62. The most common combination of item components used as input was item stem, correct answer, and options, reported 19 times (30.65%) followed by item stem only reported nine times (14.52%), and item stem and correct answer reported six times (9.68%).

Some articles from the language proficiency tests included reading passages in the input; item stem, reading passage, correct answer and options was reported nine times (14.52%), and item stem and reading passage was reported seven times (11.29%). In general, utilizing all item components appears to be the most frequently used input text source for item difficulty modeling in the reviewed studies.

4.8 Features

A total count of 131 feature groups were found in 46 studies, which are generally categorized as hand-crafted features or embeddings. This can be further classified into five broad categories: hand-crafted linguistic features, features related to item metadata, LLM generated features, static embeddings, and contextualized embeddings.

The first category of hand-crafted features are linguistic features (79 counts, 60.31%), and they consist of lexical features (e.g., number of words, length of words), syntactic features (e.g., sentence count, use of conjunctions), morphological features (e.g., word stems, lemmas), semantic features (e.g., semantic similarity between item stem and options), readability indices (e.g., Flesch Reading Ease, Gunning FOG Index), and content specific features (e.g., number of text-based numerical

[1] RACE++ is a large-scale reading comprehension dataset; IFLYTEK refers to a language dataset from iFlytek, a Chinese technology company.

values for math). The second set of hand-crafted features include item metadata features (21 counts, 16.03%) including cognitive complexity, content standards, expert ratings, and item characteristics (e.g., number of choices for MC items). The third type of hand-crafted features are reasoning and thinking level features generated from LLMs (5 counts, 3.82%). Some examples of this type include first-token probability, choice-order sensitivity, and justification length.

As for embedding features, there are two categories: static embeddings and contextualized embeddings. Static embeddings (8 counts, 6.11%) include count-based model embeddings (e.g., Glove), and predictive model embeddings (e.g., word2vec). Contextual embeddings (18 counts, 13.74%) include deep learning-based embeddings (e.g., ELMo), word-level SLM embeddings (e.g., BERT-base, DistilBERT, MPNet), sentence-level SLM embeddings from sentence-BERT and Longformer, and embeddings extracted from LLMs.

4.9 Models

Among all reviewed studies, a total of 61 models have been explored 160 times, as it was common for studies to compare multiple models. The models were classified into three categories: classical machine learning models (94 counts, 58.75%), neural network based deep learning models (20 counts, 12.50%), and transformer-based language models (46 counts, 28.75%). Among the transformer models, 39 counts (84.78%) were SLMs and 7 (15.22%) were LLMs. For a full list of models included in the review see Appendix A.

Classical machine learning models typically rely on engineered features and are built on either statistical assumptions or algorithmic decision rules. They can be further classified as follows: linear and penalized regression models, decision tree-based models, probabilistic models, ensemble learning methods, kernel and distance-based models, and simple neural network-based models.

Neural network-based deep learning models use multiple layers that mimic the functioning of human neurons to learn complex, non-linear representations from data. In this review, we define this category as including only neural network models with more than one hidden layer and that are not based on attention mechanisms. These models consisted of basic neural network architectures, convolutional neural networks, bi-directional long short-term memory, and embeddings from language models (ELMo).

Transformer-based language models represent a specialized subset of deep learning in which the transformer architecture, characterized by self-attention mechanisms, is employed. The self-attention mechanism contextualizes each word in the text by considering its relationship with all other words, regardless of position or distance. This category contains both SLMs and LLMs, where we defined SLM as language models containing less than 1 billion parameters. SLMs consisted of BERT and its variants, long-sequence transformers, T5, and GPT-2. LLMs consisted of the models in the families of GPT, Llama, Mistral-7B, Gemma-7B, Qwen-2, Yi-34b, and Phi3, though Claude and Gemini families could be utilized as well.

We note several trends about the use of models through the years. Classical machine learning techniques have retained momentum due to their transparency, interpretability, efficiency, and robustness with small sample sizes. Neural network based deep learning models have been intermittently used beginning in 1995 and gaining moderate traction in 2019 to 2020. During this time, the use of neural-network-based deep learning models was approximately equal to the use of classical machine learning models. However, there has been a decline in the use of neural network based deep learning models that coincides with the rise of transformer-based models around 2020. Since then, classical machine learning models are still used, while transformer-based models have been used for both predicting item difficulty and generating embeddings as features.

4.10 Evaluation Criteria

Model performance was assessed using 23 unique evaluation criteria and their application depended on whether item difficulty prediction was a regression or classification task. In this review 43 studies were regression tasks and 3 were classification tasks. With regards to the regression tasks, the most common evaluation criteria were root mean square error (RMSE) (k=28, 31.82%), Pearson product moment correlation (k=17, 19.32%), and R^2 (k=13, 14.77%), mean absolute error (k=8, 9.09%), and mean square error (k=5, 5.68%). For classification tasks, exact accuracy was used for each study (k=3, 37.50%), and

adjacent accuracy was used when there were more than three difficulty levels (k=2, 25.00%). F1-score, recall, and precision were used once (12.50% each).

4.11 Model Performance

Appendix B presents a table that summarizes the best-performing value for each evaluation criteria used in the reviewed studies. Evaluation criteria that were used twice were summarized with the minimum and maximum values. It is important to note that although the table summarizes the outcomes from most commonly used evaluation criteria, values are not directly comparable across studies due to different subject domains and item difficulty parameters without a common scale. Instead, it should be used to provide a sense of what constitutes a "typical result" based on the range and distribution of values obtained in the literature.

For the most commonly used evaluation metric in regression tasks, RMSE, the summary was made for p-value, transformed p-value, and Rasch model b-parameter. The RMSE for studies using p-value ranged from 0.165 to 0.268 (N=6, M=0.216, SD=0.035), while RMSE for studies using transformed p-values ranged from 0.253 to 0.308 (N=10, M=0.291, SD=0.018). RMSE for studies using the Rasch model b parameter ranged from 0.354 to 1.295 (N=8, M = 0.740, SD = 0.297). The RMSE based on other difficulty parameters (i.e., 3PL, Delta, categorical levels), only contained one value for each, therefore a meaningful summary could not be produced.

The pattern persisted for other regression evaluation metrics. The Pearson correlation ranged from 0.040 to 0.870 (N=17, M=0.545, SD=0.225). R^2 values ranged from 0.208 to 0.788 (N=13, M=0.478, SD=0.200).

Similarly, the range for classification evaluation metrics also varied greatly across studies. Exact accuracy ranges from 0.325 to 0.806 (N=3, M=0.567, SD=0.241). However, the moderate to very high adjacent accuracy values (N = 2, 0.65 and 0.982) indicate that even when the model's prediction is not exactly correct, it is close, often just one category away from the true difficulty level.

5 Discussion

The aim of this review was to highlight and summarize trends in text-based item difficulty prediction research in the large-scale assessment setting, with a focus on advanced machine learning and language model-based approaches. A total of 46 studies from 37 articles were synthesized and results showed high potential for automated prediction of item difficulty parameters.

Our review makes several contributions to large-scale educational assessments. We provide large-scale educational assessment programs with foundational information that can be used to guide the implementation of automated approaches for item difficulty in the test development process. We provide practical insights into the optimal input and prompting strategies such as including all item components in the input and using a larger portion of the data for training leads to increased model performance. Our review can also be used as guidance for model and feature selection, outlining critical considerations for methodological choices. Overall, automated item difficulty modeling can be used to reduce the time and cost of traditional field testing to evaluate item quality.

Additionally, this review presents major contributions to the field of machine learning. Unlike previous reviews that have only synthesized the literature through 2022, the present review captures the significant growth of research in the past three years, as well as how methodological approaches have evolved since the 1990s. Another unique contribution of our review is the numerical distribution of model performance outcomes across all studies. The distribution of outcomes acts as a reference point that future researchers can use to set realistic expectations and to contextualize their model performance results.

Nonetheless, our review has a few limitations including the potential bias due to the overrepresentation of papers from the BEA shared task, lack of diversity in certain aspects of the datasets (e.g., item type, content domain), unexplained variability in model performance, and limited reporting of observed range of the IRT b-parameter. The latter complicates interpretation of scale-dependent evaluation metrics that were summarized in the model performance section. Future studies should prioritize dataset diversity, transparent reporting of methodology, and approaches that balance interpretability with the capabilities of state-of-the-art language models.

References

References marked with an asterisk (*) indicate studies included in the meta-analysis.

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. International Journal of Artificial Intelligence in Education, 34(3), 862-914.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

*Aryadoust, V. (2013, April). Predicting item difficulty in a language test with an Adaptive Neuro Fuzzy Inference System. In 2013 ieee workshop on hybrid intelligent models and applications (hima) (pp. 43-50). IEEE.

*Aryadoust, V., & Goh, C. C. (2014). Predicting listening item difficulty with language complexity measures: A comparative data mining study. CaMLA Work. Pap, 2, 1-16.

*Beinborn, L., Zesch, T., & Gurevych, I. (2015). Candidate evaluation strategies for improved difficulty prediction of language tests. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1–11). Denver, CO: Association for Computational Linguistics.

*Benedetto, L. (2023, June). A quantitative study of NLP approaches to question difficulty estimation. In International Conference on Artificial Intelligence in Education (pp. 428-434). Cham: Springer Nature Switzerland.

*Boldt, R. F., & Freedle, R. (1996). Using a neural net to predict item difficulty. *ETS Research Report Series*, *1996*(2), i-19.

*Boldt, R. F. (1998). GRE analytical reasoning item statistics prediction study. *ETS Research Report Series*, *1998*(2), i-23.

*Bulut, O., Gorgun, G., & Tan, B. (2024). Item Difficulty and Response Time Prediction with Large Language Models: An Empirical Analysis of USMLE Items.

*Dueñas, G., Jimenez, S., & Ferro, G. M. (2024, June). Upn-icc at bea 2024 shared task: Leveraging llms for multiple-choice questions difficulty prediction. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 542-550).

*El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J. A. (2016). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal*, 28(1), 59–82.

*Feng, W., Tran, P., Sireci, S., & Lan, A. (2025). Reasoning and Sampling-Augmented MCQ Difficulty Prediction via LLMs. arXiv preprint arXiv:2503.08551.

Ferrara, S., Steedle, J. T., & Frantz, R. S. (2022). Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education*, 35(3), 237-253.

*Fulari, R., & Rusert, J. (2024, June). Utilizing Machine Learning to Predict Question Difficulty and Response Time for Enhanced Test Construction. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 528-533).

*Gombert, S., Menzel, L., Di Mitri, D., & Drachsler, H. (2024, June). 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 (BEA 2024) (pp. 483-492).

*Groot, N. (2023). *Using Task Features to Predict Item Difficulty and Item Discrimination in 3F Dutch Reading Comprehension Exams* (Master's thesis, University of Twente).

*Ha, L. A., Yaneva, V., Baldwin, P., & Mee, J. (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 (pp. 11–20). Florence, Italy: Association for Computational Linguistics.

*He, J., Peng, L., Sun, B., Yu, L., & Zhang, Y. (2021). Automatically predict question difficulty for reading comprehension exercises. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1398-1402).

*Hsu, F. Y., Lee, H. M., Chang, T. H., & Sung, Y. T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. Information Processing & Management, 54(6), 969–984.

*Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., ... & Hu, G. (2017, February). Question Difficulty Prediction for READING Problems in Standard Tests. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).

*Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., & Domingue, B. W. (2025). Prediction of Item Difficulty for Reading Comprehension Items by Creation of Annotated Item Repository. arXiv preprint arXiv:2502.20663.

Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., ... Dowling, W. (2021). A comparative study of methods for a priori prediction of MCQ difficulty. Semantic Web, 12(3), 449–465

*Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., Lissitz, R. (2025). Item Difficulty Modeling Using Fine-Tuned Small and Large Language Models. Educational and Psychological Measurement. (accepted).

Luecht, R. M. (2025). Assessment engineering in test design: Methods and applications. Taylor & Francis.

*Loukina, A., Yoon, S. Y., Sakano, J., Wei, Y., & Sheehan, K. (2016, December). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3245-3253).

*McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-Starting Item Parameters for Adaptive Language Tests. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Pp. 883-899.

*Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. Language Testing, 12(1), 34-53.

*Qunbar, S. A. (2019). Automated item difficulty modeling with test item representations (ERIC No. ED601723). [Doctoral dissertation, The University of North Carolina at Greensboro].

*Razavi, P., & Powers, S. J. (2025). Estimating Item Difficulty Using Large Language Models and Tree-Based Machine Learning Algorithms. arXiv preprint arXiv:2504.08804.

*Rodrigo, A., Moreno-Álvarez, S., & Peñas, A. (2024, June). Uned team at bea 2024 shared task: Testing different input formats for predicting item difficulty and response time in medical exams. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 567-570).

*Rogoz, A. C., & Ionescu, R. T. (2024). UnibucLLM: Harnessing LLMs for Automated Prediction of Item Difficulty and Response Time for Multiple-Choice Questions. *arXiv preprint arXiv:2404.13343*.

*Sano, M. (2015). Automated capturing of psycho-linguistic features in reading assessment text. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

SIGEDU. (2024). BEA 2024 Shared Task: Automated Prediction of Item Difficulty and Item Response Time. https://sig-edu.org/sharedtask/2024

*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), 1-30.

*Tack, A., Buseyne, S., Chen, C., D'hondt, R., De Vrindt, M., Gharahighehi, A., Metwaly, S., Nakano, F. K., & Noreillie, A.-S. (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) (pp. 512–521).

*Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. Language Testing, 34(2), 151–174.

*Veeramani, H., Thapa, S., Shankar, N. B., & Alwan, A. (2024, June). Large Language Model-based Pipeline for Item Difficulty and Response Time Estimation for Educational Assessments. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 561-566).

*Xue, K., Yaneva, V., Runyon, C., & Baldwin, P. (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.

*Xue, M., Han, S., Boykin, A., & Rijmen, F. (2025, April). *Leveraging Large Language Models in Predicting Item Difficulty*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

*Yaneva, V., Baldwin, P., & Mee, J. (2019, August). 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* (pp. 11-20).

*Yaneva, V., Ha, L. A., Baldwin, P., & Mee, J. (2020). Predicting item survival for multiple-choice questions in a high-stakes medical exam. *In Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6812–6818). European Language Resources Association.

*Yi, X., Sun, J., & Wu, X. (2024). Novel feature-based difficulty prediction method for mathematics items

using XGBoost-based SHAP model. Mathematics, 12(10), 1455.

*Yousefpoori-Naeim, M., Zargari, S., & Hatami, Z. (2024, June). 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 (BEA 2024)* (pp. 551-560).

*Zotos, L., van Rijn, H., & Nissim, M. (2024). Are You Doubtful? Oh, It Might Be Difficult Then! Exploring the Use of Model Uncertainty for Question Difficulty Estimation. arXiv preprint arXiv:2412.1183.

Appendix A. List of Models Included in the Review.

Classical Machine Learning Models:

1. *Linear and Penalized Regression Models:* Ordinary Least Square Regression, Principal Components Regression, Partial Least Squares Regression, Elastic Net Regression, Lasso Regression, Ridge Regression/Ridge (L2) Penalized Regression, Linear Logistic Test Model (LLTM)
2. *Decision Tree-Based Models:* Classification and Regression Trees (CART), Classification Trees, Decision Tree Regression, Extra Trees, Random Forest, Regression Trees
3. Probabilistic Models: Naive Bayes Classifier, Gaussian Processes, Probabilistic language model
4. *Ensemble Learning Models*: AdaBoost, Cat-Boost, Gradient Boosting, Gradient Boosting Decision Trees, Light Gradient Boosting Machine, XGBoost, XGBoost-based SHAP Model
5. *Kernel and Distance-Based Models*: k-Nearest Neighbors, Support Vector Machines
6. *Simple Neural Network Based Models*: Adaptive Neuro-Fuzzy Inference System (ANFIS), One Neuron Network (with no hidden layer), Three-Layer Backpropagation Neural Network (with only one hidden layer)

Neural Network Based Deep Learning Models:

1. *Basic Neural Network Architectures*: Artificial Neural Network (ANN), Multilayer-Perceptron (MLP), Dense Neural Network
2. *Convolutional Neural Networks (CNNs) and Variants*: Convolutional Neural Network (CNN), Attention-based CNN (ACNN), Hierarchical Attention-Based CNN (HBCNN), Multi-Scale Attention CNN (MACNN), Temporal CNN (TCNN), Temporal Attention CNN (TACNN).
3. *Bidirectional Long Short-Term Memory (Bi-LSTM)*

Transformer-Based Language Models
Small Language Models:

1. *BERT and its Variants*: BERT, BERT-ClinicalQA, Clinical-BERT, BioClinicalBERT, Bio_ClinicalBERT_emrqa, Bio_ClinicalBERT_FTMT, Clinical-BigBird, BioMedBERT, PubMedBERT, DistilBERT, ConvBERT, DeBERTa, RoBERTa, Electra, BioMedElectra
2. *Long-Sequence Transformers:* Longformer, Clinical-Longformer, Longformer-Base-4096, BigBird
3. *GPT-2*
4. *T5*

Large Language Models:

1. *GPT Family*: GPT-4, GPT-4o
2. Llama-7B
3. *Mistral*-7B
4. *Gemma*-7B
5. *Phi 3*

Appendix B. Model Performance Summary.

Evaluation Criterion	Count	Min	Max	Median	Mean	SD
Regression Tasks						
RMSE						
Based on p-value	6	.165	.268	.214	.216	.035
Based on transformed p-value	10	.253	.308	.297	.291	.018
Based on Rasch model	8	.354	1.295	.693	.740	.297
MSE	5	.013	.521	.064	.203	.227
MAE	7	.185	.58	.240	.307	.159
Correlation						
Pearson	17	.04	.87	.550	.545	.225
Spearman	4	.25	.790	.496	.508	.221
R-Squared	13	.208	.788	.525	.478	.200
Match	2	.757	.780	-	-	-
Classification Tasks						
Accuracy						
Exact	3	.325	.806	.569	.567	.241
Adjacent	2	.65	.982	-	-	-

Note. For studies that reported multiple models or evaluation criteria, only the best-performing value for each evaluation criterion was included. Only evaluation criteria that provided enough information ($k \geq 2$) for meaningful analysis were included. We also note that we report the same number of decimals that were presented in the articles.