Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions

Victoria Yaneva[1], Kai North[2], Peter Baldwin[1], Le An Ha[3], Saed Rezayi[1],
Yiyun Zhou[1], Sagnik Ray Choudhury[1], Polina Harik[1], and Brian Clauser[1]

[1]National Board of Medical Examiners, Philadelphia, USA
{vyaneva, pbaldwin, srezayidemne, yyzhou, sraychoudhury, pharik, bclauser}@nbme.org
[2]George Mason University, USA
knorth8@gmu.edu
[3]Ho Chi Minh City University of Foreign Languages, Vietnam
anhl@huflit.edu.vn

Abstract

This paper reports findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions. The task was organized as part of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA'24), held in conjunction with NAACL 2024, and called upon the research community to contribute solutions to the problem of modeling difficulty and response time for clinical multiple-choice questions (MCQs). A set of 667 previously used and now retired MCQs from the United States Medical Licensing Examination (USMLE®) and their corresponding difficulties and mean response times were made available for experimentation. A total of 17 teams submitted solutions and 12 teams submitted system report papers describing their approaches. This paper summarizes the findings from the shared task and analyzes the main approaches proposed by the participants.

1 Introduction

For standardized exams to be fair and defensible, test items must meet certain criteria. One important criterion for many exams is that the questions cover a wide range of difficulty levels to allow information about a wide range of examinee proficiencies to be collected effectively. Additionally, it is often essential to allocate an appropriate amount of time for each question: too little time can make the exam speeded, while too much can make it inefficient. Often, item difficulty and response time data are collected via a process called *pretesting*, wherein new items appear on live exams alongside scored items. While robust, the need for a statistically sufficient sample of examinees to complete these items restricts the number of items that can be pretested, potentially leading to overexposure and jeopardizing item security (Settles et al., 2020).

The problem of estimating item characteristics with little to no response data is a decades-old research topic. Early studies used what is sometimes referred to as auxiliary or collateral information—including various properties of an item's text—to improve parameter estimation within a Bayesian framework (Mislevy, 1988; Stowe, 2002; Swaminathan et al., 2003). Recent advances in NLP have led to a renewed interest in predicting item characteristics based on item text. As with the earlier research, it is hoped that such predictions may be used to "jump-start" parameter estimation (McCarthy et al., 2021) allowing items to be exposed to fewer test-takers, or improve fairness by making the time intensiveness of test forms that include pretest items less variable (Baldwin et al., 2020).

While there is evidence that NLP techniques may offer a potential solution (see Section 2), the absence of publicly available datasets has resulted in fragmented efforts to advance the state-of-the-art in item parameter prediction, impeding meaningful comparisons between different approaches, exacerbating issues of reproducibility, and stifling collaboration. Furthermore, as outlined in Section 2, the existing literature has concentrated on difficulty prediction, neglecting other crucial item parameters such as response time, which also have important implications for exam fairness and validity.

To address these shortcomings and advance this area of research, we organized the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple Choice Questions[1]. The shared task was organized as part of the 19th

[1] https://sig-edu.org/sharedtask/2024

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 470–482
June 20, 2024 ©2024 Association for Computational Linguistics

Workshop on Innovative Use of NLP for Building Educational Applications (BEA'24), collocated with NAACL 2024, and took place between January 15 and March 10, 2024. An ideal shared dataset for this task would encompass test items along with their corresponding difficulties and response times based on responses collected from a sufficiently large and diverse examinee sample under standardized test conditions. To this end, 667 retired clinical multiple-choice questions (MCQs) from a high-stakes medical exam[2] were released for the exploration of two topics: predicting item difficulty (Track 1) and predicting item response time (Track 2). Overall, 48 teams enrolled as participants, of which 17 submitted solutions and 12 submitted system review papers describing their approaches. This paper summarizes the organization and main findings from the competition. The data are available upon request at https://www.nbme.org/services/data-sharing.

2 Related Work

This section summarizes the main approaches used in item difficulty and response time prediction research, with special emphasis on clinical MCQs, the domain of the shared task. For a systematic review of the literature, we refer the reader to AlKhuzaey et al. (2023).

2.1 Predicting Item Difficulty

Most of the early research on modeling item difficulty was in the domain of language learning and used predictors such as lexical, syntactic, statistical, and readability features. Freedle and Kostin (1993) and Perkins et al. (1995) used a mix of lexical and syntactic features, such as vocabulary, sentence and paragraph length, number of negations and referentials, and lexical overlap between text and options to determine the difficulty of MCQs from English foreign language exams and reading comprehension tests, respectively. These features were later expanded to cohesion, discourse, and psycholinguistic features among others (Beinborn et al., 2014, 2015; Loukina et al., 2016).

Outside the domain of language learning, these features showed comparatively weaker predictive power. El Masri et al. (2017) found that linguistic features were not good predictors for item difficulty in middle-school science items, "likely due to the extent to which computational linguistic facilities are less effective with very short textual materials". Likewise, Susanti et al. (2017) and Benedetto et al. (2020) found that readability metrics were relatively poor predictors of item difficulty for computer science and English vocabulary MCQs, respectively.

Consistent with other NLP use cases, more recent studies on item parameter prediction utilize neural approaches. Huang et al. (2017) used embeddings and an attention-based convolutional neural network to predict the difficulty of reading items. Hsu et al. (2018) converted items into word-embeddings, calculated the cosine similarities between stem, answer, and distractors, and used them to train a support vector machine (SVM) to predict item difficulty of MCQs from the domain of social studies. Zhou and Tao (2020)'s fine-tuned BERT model (Devlin et al., 2018) achieved a higher F1-score for predicting item difficulty of open-ended programming-related questions compared to a Bidirectional Long Short-Term Memory (BiLSTM) model. Benedetto et al. (2021) trained a series of BERT and DistilBERT models with several pre-training steps, including the use of masked-language modeling. BERT achieved the highest performance for predicting item difficulty of math and computer science open-ended questions and MCQs, having surpassed all other models—including several word-embedding approaches. Other notable studies in this area include Loginova et al. (2021) and He et al. (2021).

Item difficulty prediction has also been applied in efforts to automatically generate items at desired levels of difficulty (e.g., Gao et al. (2018), Bi et al. (2021)). Some of these approaches assess the semantic similarity between a question and its associated answer choices (Alsubait et al., 2013; Kurdi et al., 2016), while others focus on items that assess an examinee's ability to distinguish between words and pseudo-words, and thus utilize word and sub-word level predictors (Settles et al., 2020).

2.2 Predicting Item Response Time

The prediction of response time is a less-researched area, further motivating its inclusion within this shared task. Early studies included features such as the sequential position of the item within an exam (Parshall et al., 1994), the inclusion of visual aids (Smith, 2000; Swanson et al., 2001), and word-count (Halkitis et al., 1996; Smith, 2000).

[2] The United States Clinical Licensing Examination (USMLE®)

A 65-year-old woman comes to the physician for a follow-up examination after blood pressure measurements were 175/105 mm Hg and 185/110 mm Hg 1 and 3 weeks ago, respectively. She has well-controlled type 2 diabetes mellitus. Her blood pressure now is 175/110 mm Hg. Physical examination shows no other abnormalities. Antihypertensive therapy is started, but her blood pressure remains elevated at her next visit 3 weeks later. Laboratory studies show increased plasma renin activity; the erythrocyte sedimentation rate and serum electrolytes are within the reference ranges. Angiography shows a high-grade stenosis of the proximal right renal artery; the left renal artery appears normal.
Which of the following is the most likely diagnosis?
(A) Atherosclerosis
(B) Congenital renal artery hypoplasia
(C) Fibromuscular dysplasia
(D) Takayasu arteritis
(E) Temporal arteritis

Table 1: An example of a practice item from the USMLE Step 1 Sample Test Questions (`usmle.org`). © 2024 National Board of Medical Examiners and the Federation of State Medical Boards, used with permission.

Schneiderand et al. (2023) is one of the few studies that used text-based features to predict student response time for items on multiple topics, ranging from everyday life to personality and politics. They trained models such as stochastic gradient boosting (SGB), SVM, and random forests (RF) on 51 features including question length, lexical diversity, and readability features, such as number of complex words, with SGB achieving best performance.

2.3 Focus on Clinical MCQs

The studies most relevant to this shared task are the ones focused on predicting characteristics of clinical MCQs from the USMLE exam. These include Ha et al. (2019), who used a 113 linguistic features and different embedding types to predict the difficulty (proportion correct responses) of 12,038 items. This study indicated that predicting item difficulty for this domain is a challenging task, with Root Mean Squared Error (RMSE) of .225 for the best result compared to a dummy regressor baseline of .237. Baldwin et al. (2020) built upon this study by applying the same predictors to the problem of modeling response time, and showed that exam fairness can be improved through meaningful reductions in the variability of time intensiveness across test forms when predicted response times for pretest items are taken into accounted during form assembly. Xue et al. (2020) applied transfer learning to the prediction of item parameters and showed that the prediction of difficulty can be improved by incorporating response time during training, but not vice-versa. Yaneva et al. (2020) aimed to automatically identify items that meet statistical criteria for live use in terms of both difficulty and discrimination[3]. Yaneva et al. (2021) examined the relationship between the linguistic characteristics of a test item and the complexity of the response process required to answer it correctly, defined as the interaction between difficulty and response time. The methods used in these studies are summarized in Yaneva et al. (2023), which was written for educational measurement professionals and provides an overview of the applications of NLP methods to this task.

3 Shared Task Description

The data for the shared task comprises 667 previously used and now retired MCQs from Steps 1, 2 CK, and 3 of the United States Medical Licensing Examination (USMLE®). USMLE is a sequence of examinations (called *Steps*), developed by the National Board of Medical Examiners (NBME®) and Federation of State Medical Boards (FSMB), that is used to support medical licensure decisions in the United States. Each step includes 7 to 12 blocks of MCQ items (a block ranges between 45 and 60 minutes), and each item is answered by approximately 300+ examinees. Item characteristics used in this shared task were based on examinees who were medical students from accredited[4] US and Canadian medical schools taking the exam for the first time.

An example practice item from the dataset is given in Table 1. The part describing the case is referred to as *stem*, the correct answer is referred to as *key*, and the incorrect answer options are known as *distractors*. All items test medical knowledge

[3] Item discrimination is a measure of the extent to which an item differentiates between students of different proficiency.

[4] Accredited by the Liaison Committee on Medical Education (LCME).

and were written by experienced subject-matter experts following a set of guidelines. These guidelines stipulate adherence to a standard structure, as well as the avoidance of extraneous material not needed to answer the item, information misleading the test-taker, or correct answers that are longer or more specific than the other options.

Each item is tagged with metadata indicating whether or not it contains an image, the Step exam it was presented on, as well as Difficulty and Response Time data, as shown in the structure below:

- *ItemNum* denotes the consecutive number of the item in the dataset (e.g., 1,2,3,4,5, etc).
- *ItemStem_Text*: the text of the item stem (the part of the item describing the clinical case).
- *Answer_A*: the text for response option A
- *Answer_B*: the text for response option B
- *(. . .)*
- *Answer_J*: the text for response option J. For items with fewer than J response options, the remaining columns are left blank. For example, if an item contains response options A to E, the fields for columns F to J are left blank for that item.
- *Answer_Key*: the letter of the correct answer.
- *Answer_Text*: the text of the correct answer.
- *ItemType*: whether the item contained an image (e.g., an x-ray image, picture of a skin lesion, etc.) or not. The value "Text" denotes text-only items and the value "PIX" denotes items that contain an image. Note that the images are not part of the dataset.
- *EXAM*: The USMLE Step (1, 2 or 3) the item was presented on. For more information on the Steps of the USMLE see `https://www.usmle.org/step-exams`.
- *Difficulty*: The (linearly-transformed) proportion of correct responses across all examinees who attempted a given item during a live exam. After the transformation, higher values indicated more difficult items.
- *Response_Time*: arithmetic mean response time, measured in seconds, across all examinees who attempted a given item on a live exam. This includes all time spent on the item from the moment it is presented on the screen until the examinee moves to the next item, as well as any time spent revisiting the item.

The task was divided in two tracks as follows:

- Track 1: Given the item text and metadata, predict the item difficulty variable.
- Track 2: Given the item text and metadata, predict the time intensity variable.

Out of the full sample, 466 items were made available as a labeled training set and the other 201 items were retained as an evaluation set. Training data outside of the specified training set were allowed, provided these data were publicly available and their license allows use for research purposes. Use of one target variable in the prediction of another was *not* permitted, since in most cases, predicting these variables will be most beneficial prior to the collection of response data—at which time neither the difficulty nor the time intensity parameters can be estimated.

Submissions were requested as separate .csv files for each track. Each file had to contain the item number (Item_Num) and corresponding predicted value for each item. Teams were allowed to submit up to three attempts per track, differentiated by adding run1, run2, or run3 to the name of their uploaded .csv file; however, such teams were required to explain how each attempt differed within their system report paper—i.e., changes in methodology, parameters, models used, prediction strategy, etc.

In both tracks, the evaluation was based on RMSE, and teams that achieved the lowest RMSE value were considered winners. There were two separate leaderboards for Track 1 and Track 2. In both, submissions were ranked according to the RMSE metric from Python's scikit-learn library (Pedregosa et al., 2011).

4 Results

A total of 17 teams submitted up to 3 solutions for item difficulty prediction and 15 teams submitted up to 3 solutions for response time prediction. Table 2 presents ranked results for the top 15 solutions in both tracks. The full leaderboard is available at `https://sig-edu.org/sharedtask/2024#results`.

In Track 1, Predicting Item Difficulty, there are minor differences between the RMSE of the top 15 solutions; however, even the best solution outperformed the baseline by only a small margin (#1, EduTec = 0.299, #16, DummyRegressor = 0.311). These results are consistent with the prior literature

Difficulty				Response Time			
Rank	Team Name	Run	RMSE	Rank	Team Name	Run	RMSE
1	EduTec	electra	0.299	1	UNED	run2	23.927
2	UPN-ICC	run1	0.303	2	ITEC	Lasso	24.116
3	EduTec	roberta	0.304	3	UNED	run1	24.777
4	ITEC	RandomForest	0.305	4	UNED	run3	25.365
5	BC	ENSEMBLE	0.305	5	EduTec	roberta	25.64
6	Scalar	Predictions	0.305	6	EduTec	electra	25.875
7	BC	FEAT	0.305	7	UnibucLLM	run3	26.073
8	BC	ROBERTA	0.306	8	ED	run1	26.57
9	UnibucLLM	run1	0.308	9	Rishikesh	1	26.651
10	EDU	Run3	0.308	10	UnibucLLM	run2	26.768
11	EDU	Run1	0.308	11	UnibucLLM	run1	26.846
12	ITEC	Ensemble	0.308	12	SCaLARlab	run3	26.945
13	UNED	run3	0.308	13	Scalar	predictions	26.982
14	Rishikesh	1	0.31	14	EduTec	deberta	27.302
15	Iran-Canada	run2	0.311	15	EDU	Run1	27.474
16	Dummy Regressor Baseline		0.311	25	Dummy Regressor Baseline		31.68

Table 2: Top 15 leaderboard results for Track 1: Difficulty and Track 2: Response Time

on clinical MCQs presented in Section 2.3, underscoring the challenging nature of the task. In Track 2, Response Time, the solutions are relatively more successful compared to the DummyRegressor baseline (#25 DummyRegressor, RMSE = 31.68), with the #1 solution obtaining RMSE of 23.927.

Of the 17 teams who submitted solutions, 12 submitted system report papers, which are summarized below (10 papers for both Track 1 and Track 2, and 2 papers only for Track 1).

5 Main Approaches

The solutions submitted by the participants encompassed several approaches that had not been previously applied to the problem of modeling item characteristics. Some of these were comparatively simpler models that performed unexpectedly well, such as the case of the submission that ranked #1 in predicting response time (Rodrigo et al., 2024). In the case of modeling item difficulty, several approaches used classical methods such as linguistic features combined with embeddings but expanded the set of features to include novel predictors. These traditional solutions were not as successful for item difficulty prediction, which favored more novel approaches. These novel approaches can be broadly categorized as transformer model modifications, question answering using LLMs, and data augmentation techniques. These categories are not necessarily mutually exclusive (e.g., some approaches use both data augmentation and linguistic features); however, we found this broad classification scheme useful in describing the submitted solutions, as shown below. The main techniques used in the studies are further summarized in Section 5.6.

5.1 Efficient solutions that performed well

Well-performing solutions include the ones proposed by **UNED** (Rodrigo et al., 2024), who focused on feeding combinations of the full item, stem and correct answer, or stem only into a BERT base model (Devlin et al., 2018). The three submissions differed only by these input configurations and were the same for both tracks (with different target variables). There was no special preprocessing and the tokenzier was the one provided by the BERT model. The target variables were both scaled [0-1]. Perhaps somewhat surprisingly given its simplicity, this system ranked #1 for response time prediction (RMSE of 23.927 with text and correct answer as input) and #13 for difficulty prediction (RMSE of 0.308, stem only).

Scalar (DataWizards) concatenated BERT embeddings with TF-IDF encodings for item difficulty prediction and Word2Vec embeddings with TF-IDF encodings for response time prediction. These representations of different item components (e.g., stem only or stem + answer options) were used as predictors in various models, of which RF performed best. This solution ranked #6 for predicting item difficulty (RMSE = 0.305) and #13 for response time (RMSE = 26.982).

These solutions serve as an important benchmark for the added value provided by the linguistic features, question-answering techniques, and model optimization approaches presented next.

5.2 Transformer model modifications

The solution that ranked #1 for predicting difficulty was from the category of novel model optimization techniques. **EduTec** (Gombert et al., 2024) proposed optimizing pre-trained transformer encoder language models using three modifications. The first modification was the use of scalar mixing, which is a procedure that calculates a weighted mean of all hidden layers of the transformer (the weights are fit during training). Scalar mixing is hypothesized to be helpful because, as different layers within transformer models learn representations for different linguistic phenomena, it allows the use of representations from all these different layers (as opposed to the final layer alone), while simultaneously learning their importance for the final output. The second modification was a two-layer setup for the classification heads, where the input from the intermediate layer was run through a *rational activation*: a form of learnable activation function whose shape is optimized during training. This type of activation function was shown to outperform non-learnable activation functions. Third, the authors used multi-task learning to learn shared representations for both difficulty and response time, motivated by the observed correlation between the two variables within the training set. The architecture described so far was evaluated with different transformer encoder models, of which ELECTRA achieved #1 in the shared task leaderboard for difficulty prediction with an RMSE of 0.299 and #6 on the leaderboard for response time prediction (RMSE = 25.875). RoBERTa achieved #5 for response time prediction with an RMSE of 25.64.

5.3 Question answering using LLMs

Two teams used responses from LLMs to extract predictive features or perform data augmentation.

UPN-ICC (Dueñas et al., 2024) investigated the hypothesis that item difficulty depends more on the features of the test-taking population than on the items themselves. They simulated medical students' answers to the MCQs by prompting chatGPT 3.5 in four different settings: i) answering each question and providing a brief justification for the response, ii) providing a yes/no response for each answer option on whether it is the correct answer, iii) randomly removing 20% of the content tokens from the stem to simulate examinees who did not read the item carefully, and iv) all of the above but with a varying temperature parameter[5]. The justification behind iv) is the hypothesis that items that are only answered correctly under a low temperature condition can be considered difficult, while items answered correctly under any temperature can be considered easier. Next, the authors extracted more than 40 features from the generated output of the question-answering experiments. Examples of such features include "A Boolean indicating whether or not the question was answered correctly by the LLM" and "Time in milliseconds reported by the LLM to answer the question" for condition i), "Number of sub-items answered correctly for the item" for condition ii), "Boolean indicating if the LLM answered correctly the question in spite of the stem being mutilated at 20% of its content words (other six features for 30%, 40%, 50%, 60%, 70%, and 80%" for condition iii), and "Number of incorrect answers for the item out of the 11 values of *t* [temperature] used" for condition iv). These features were used as input for a Ridge regression model, which ranked #2 in difficulty prediction (RMSE = 0.303). While the indicator of whether the question was answered correctly emerged as the most significant feature, all four strategies produced meaningful predictors.

UnibucLLM (Rogoz and Ionescu, 2024) hypothesized that the number of LLMs that answer an item correctly can be an indicator of its difficulty. In a zero-shot setup, they obtained responses from three LLMs (Falcon-7B, (Almazrouei et al., 2023), Meditron-7B (Chen et al., 2023), and Mistral-7B (Jiang et al., 2023)). They then created variations of the input that included the item text only or the item text together with the LLM responses. This input was used to finetune a pretrained BERT model and a pretrained GPT-2 model (Radford et al., 2019). The best solution for difficulty prediction was the BERT model finetuned over the item text + the answer text + the LLM-generated answers, which placed #9 with an RMSE of 0.308, showing a positive effect from the LLMs. For predicting response time, GPT-2 + original item text reached #7 with an RMSE of 26.073.

5.4 Data Augmentation

EDU (EduNLP) (Veeramani et al., 2024) incorporated additional data from the "Test of Narrative Language" assessment (TNL) (Fisher et al., 2019)

[5] A parameter that controls the level of randomness of the LLM output, ranging between p= 2.0 (maximum randomness) and p = 0.0 (fully deterministic).

to use in an auxiliary task. For both the shared task data and the TNL data, the authors first prompted three LLMs to annotate named entities within the data. Then, they passed each sentence with its annotated named entities as input to the LLMs, this time for the task of semantic role labeling[6]. Next, the LLMs were provided with the item, named entities, semantic roles, and the correct answer, and prompted to summarize the association between these and each answer option. The models then were instructed as follows: *"Depending on the difficulty level of the linkages between input context and [answer options], assign the input context a score in the range of 0 to 1.4"*. The best run from this approach ranked #10 for difficulty prediction (RMSE = 0.308). For modeling response time, the authors added numeric and syntactic features from LingFeat (Shaikh et al., 2022), resulting in #15 rank and an RMSE of 27.474.

SCaLARlab (Ram et al., 2024) performed data augmentation by utilizing LLMs to generate additional items with difficulty values above 0.7, to balance the training set. Three models were trained on the augmented dataset: i) BioBERT + Linguistic features as input to two different neural network architectures, ii) Word2Vec embeddings as input to various regressor models (e.g., RF, KNN, SVM), and iii) combinations of BioBERT + Linguistic features as input to the regressor models. The best run resulted in a rank of #19 for difficulty (RMSE = 0.315) and #12 for response time (RMSE = 26.945).

5.5 Linguistic features + embeddings

A number of teams experimented with combining various linguistic features with embeddings and performing model ensembling.

ITEC (Tack et al., 2024) extracted features from the Linguistic Inquiry and Word Count tool (LIWC-22) (Pennebaker et al., 2022) and TAALES 2.2 (Kyle and Crossley, 2015), which include classic linguistic features, as well as features that were not previously applied to this domain such as authenticity, clout, emotional tone, and academic vocabulary, among others. To these, the authors added Bio_ClinicalBERT embeddings (Alsentzer et al., 2019) for different combinations of item components (e.g., stem only, answer option only, etc.). These features were used as input to various regression models following feature selection and dimensionality reduction procedures. The authors also experimented with finetuning clinically pretrained BERT variations in a multi-target regression setting, as well as combining the output from all of these models into an ensemble. Best results for difficulty prediction were from RF, ranking #4 with an RMSE of 0.305, while a lasso model ranked #2 for response time prediction (RMSE = 24.116). The LWIC feature indicating the degree of "analytical thinking" for the answer options emerged as particularly noteworthy for predicting response time and, to a slightly lesser extent, difficulty.

Iran-Canada (Yousefpoori-Naeim et al., 2024) experimented with various features (including Coh-Metrix (Graesser et al., 2004) and number of medical terms) and MPNet embeddings (Song et al., 2020) as input to 15 regression models. After performing feature selection, they found that "the addition of embeddings only slightly enhances model performance", and that ensembling did not lead to major improvement. Notable features for difficulty prediction were related to cohesion, while for response time were related to length and presence of medical terms. The best run resulted in a rank of #15 for difficulty (RMSE = 0.311) and #18 for response time (RMSE = 28.714).

BC (Felice and Duran Karaoz, 2024) experimented with three approaches: i) a linear regression model using linguistic features similar to those in Ha et al. (2019), ii) several transformer models, of which RoBERTa (Liu et al., 2019) performed best, and iii) a linear regression ensemble built on the predictions of the previous two models. These systems ranked #7, #8, and #5, respectively, with an RMSE of 0.305 for the ensemble model for difficulty prediction. The BC team did not participate the response time track.

Rishikesh (Fulari and Rusert, 2024) combined embeddings from PubMedBERT-MS-MARCO (Deka et al., 2022) with linguistic features as input for a number of neural and non-neural models. The best run ranked #14 for difficulty (RMSE = 0.31) and #9 for response time (RMSE = 26.651).

BRG (Bulut et al., 2024) used Coh-Metrix features and BiomedBERT embeddings (Gu et al., 2021) within a lasso model following dimensionality reduction through PCA (Wold et al., 1987). This approach ranked #20 for predicting item difficulty (RMSE = 0.318) and #24 for response time (RMSE = 31.48).

[6] The authors also use Longformer (Beltagy et al., 2020) for named entity recognition and AllenNLP SRL (Gardner et al., 2018) for semantic role labeling.

5.6 Summary of techniques

Overall, the teams explored a wide variety of approaches, many of which performed similarly despite using different models and predictors.

Most teams experimented with all parts of the items (i.e. stem, options, correct answer), but some found different parts to be more appropriate for different tasks. The teams that used scaling were more successful, although their success cannot be solely attributed to this procedure. A variety of linguistic feature sets were explored: LWIC-22, TAALES 2.2, Coh-Metrix, SMOG, Lengths, LingFeat, as well as linguistic features from Ha et al. (2019) and Yaneva et al. (2020). The embedding types that were explored include TF-IDF, BERT, Word2Vec, Bio_ClinicalBERT, Clinical-Longformer, BERT-clinical_qa, BiomedBERT, Fastext, Bio-BERT, RoBERTa, DeBERTa, ELECTRA, MPNet, and PubMedBert-MS-MARCO. For feature engineering, the teams utilized correlation studies, multicolinearity reduction, AIC, BIC, and PCA to reduce the number of features. The modeling was performed using both traditional machine learning models (e.g., linear regression, Ridge, Lasso, ElasticNet, SGD, SVM, DT, RF, KNN, etc.) and finetuning neural models (BERT, GPT2, RoBERTa, bioBERT, XLNet, DeBERTa, DistilBERT). Customization techniques included scalar mixing, Rational Activation, multi-task learning, and a custom ANN. There was a variety of cross validation techniques: two teams used 5-folds, another two used 10-folds, and one used 5x5-fold; one team split training data into 80% and 20% training and development portions, and another split it 90% and 10% 30 times.

6 Discussion

The presented Shared Task is the first effort to benchmark the success of different methodologies on a common dataset of MCQs with known difficulties and response times. Several innovative approaches, previously unexplored in this context, were formulated. The findings are consistent with prior work, which showed that, for clinical MCQs, the prediction of item difficulty is more challenging than the prediction of response time.

6.1 Model Performance

For difficulty prediction, the models surpassed the baseline by a slight margin, with minimal variance among the solutions despite their distinct methodologies. One reason for the challenging nature of this task could be the homogeneity of the test-taker sample: the majority of questions were answered correctly by most examinees, who were highly able and motivated medical students taking the exams under high-stakes conditions as a requirement for obtaining a professional license. The models may perform differently when applied to exams targeting, for instance, K-12 students, where test-taker ability has higher variance, and difficulty distributions are more variable and less skewed. In addition, the comparable results achieved by different approaches imply multiple avenues for extracting predictive signal. An important question is whether these approaches would complement each other resulting in improved predictions.

When predicting response time, a wider variance in performance was observed, both among different models and in comparison to the baseline. A somewhat unexpected finding was the superior performance of a model solely utilizing a BERT Base model, surpassing other solutions. Another observation was the relative success of models utilizing linguistic features for predicting response time compared to their performance with predicting difficulty. Since the literature on predicting item response time is rather limited, it is not yet possible to draw inferences on how these findings compare to other exam domains.

6.2 Limitations

In formulating the shared task, we made several design choices, each contributing distinct strengths and limitations to this study.

The first decision involved utilizing proportion correct responses (known in the measurement literature as *p-values*) as the measure of item difficulty. P-values describe the interaction between an item and a sample of examinees. This sample dependency means that difficulty will only be comparable across items to the extent that the examinee samples used to calculate them are equivalent across items. (For this reason, difficulty parameters obtained using Item Response Theory (IRT) are often preferable to p-values, since they are sample independent.) A similar dependency exists for mean response time. For the data used in this shared task, examinees were randomly assigned to test forms within cohort and cohorts were reasonably stable over time making the p-values and mean response times sufficiently comparable for many expected applications.

The second design consideration was whether (and, if so, how) to rescale the target variables. Because normal distributions have many useful properties and most parametric tests make a normality assumption of one kind or another, it is not uncommon to transform data such that they approximate a normal distribution. For proportion correct, a logit transformation often accomplishes this; and for response times, a log transformation is typical. Such transformations will be familiar to researchers accustomed to working with these kinds of data and for many applications transformations like this are justified and sensible. Nevertheless, because there are other occasions when it may be preferable to keep values on their original scale, it is necessary to carefully consider an intended application for a dataset before deciding how it should be rescaled.

For example, when RMSE is used to evaluate predicted values—as it was for this shared task—nonlinear transformations have the effect of weighting errors differently depending on the values of the predictions and the target variables. Under these conditions, applying a logit transformation to proportion correct values would have the effect of weighting errors for values nearer to 1 or nearer to 0 more than the errors for values nearer to .5. While this may be desirable for certain applications, here we choose to leave the question of application open and weight all errors equally. To this end, only a linear transformation was applied to the proportion correct values and mean response times were left untransformed. Participants were, of course, free to transform the data in any manner they deemed helpful provided their predictions were submitted on the scale of the original values.

Third, the data for this task was limited to clinical MCQs, limiting the inferences that can be made about the generalizability of these methodologies to other domains. How the approaches generalize is an empirical question, however, one can speculate that they might be less effective in a math examination where items often contain minimal text, and more beneficial in reading-comprehension examinations where the text's complexity may be deliberately varied to manipulate difficulty. In an ideal world, future shared tasks on this topic should span multiple content domains and examinee populations with different characteristics, while remaining equally rigorous in terms of the conditions under which the examinee responses were collected.

6.3 Ethical Considerations

The data used in the Shared Task were obtained with the explicit permission of the data and copyright owners for the purposes of the Shared Task. Beyond this competition, the data are available upon request, following a data use agreement intended to ensure, to the extent possible, its ethical use in research. Test taker responses were used in aggregate, such that it is not possible to trace responses to individual examinees.

6.4 Impact

While benchmarking and fostering novel methodologies is a key contribution of this Shared Task, its impact reaches further. The competition spurred the development of a body of research on modeling item response time, a considerably less explored area. Moreover, many solutions were not narrowly tailored to the clinical realm and are potentially applicable to diverse domains and datasets. Further still, it is notable that the significance of these studies is not limited to the field of education—difficulty assessment beyond mere readability is an exciting frontier with implications for cognition and machine comprehension.

7 Conclusion

The First Shared Task on Automated Prediction of Difficulty and Response Time featured a set of 667 MCQs from a high-stakes clinical exam. Seventeen teams submitted solutions and twelve teams submitted system report papers. For Track 1, Item Difficulty Prediction, the best-performing solution achieved an RMSE of 0.299 compared to the DummyRegressor baseline of 0.311. For Track 2, Response Time Prediction, the best solution achieved an RMSE of 23.927 compared to 31.68 for the baseline. The paper summarized the methodologies proposed by the participants and discussed the contributions and limitations of the competition.

Despite the progress made, the challenge of predicting item characteristics remains formidable. Meeting this challenge necessitates not only the continued development of innovative methodologies but also the establishment of shared resources, such as public datasets containing reliable parameter estimates across various domains. Such efforts will facilitate cross-domain evaluation, fostering a more comprehensive understanding of the underlying mechanisms driving item difficulty and response time prediction.

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.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2013. A similarity-based theory of controlling mcq difficulty. In *e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on*, pages 283–288. IEEE.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2020. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 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*, pages 1–11.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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, Online. Association for Computational Linguistics.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. Introducing a Framework to Assess Newly Created Questions with Natural Language Processing. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I*, page 43–54, Berlin, Heidelberg. Springer-Verlag.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. *arXiv preprint arXiv:2110.06560*.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item Difficulty and Response Time Prediction with Large Language Models: An Empirical Analysis of USMLE Items. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

PRITAM Deka, ANNA Jurek-Loughrey, and P Deepak. 2022. Improved methods to aid unsupervised evidence-based fact checking for online heath news. *J. Data Intell.*, 3(4):474–504.

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*.

George Dueñas, Sergio Jimenez, and Geral Eduardo Mateus Ferro. 2024. 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)*, Mexico City, Mexico. Association for Computational Linguistics.

Yasmine H El Masri, Steve Ferrara, Peter W Foltz, and Jo-Anne Baird. 2017. Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. *The Curriculum Journal*, 28(1):59–82.

Mariano Felice and Zeynep Duran Karaoz. 2024. The British Council submission to the BEA 2024 shared task. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Evelyn L Fisher, Andrea Barton-Hulsey, Casy Walters, Rose A Sevcik, and Robin Morris. 2019. Executive functioning and narrative language in children with dyslexia. *American journal of speech-language pathology*, 28(3):1127–1138.

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

Rishikesh Fulari and Jon Rusert. 2024. Utilizing Machine Learning to Forecast Question Difficulty and Response. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Yifan Gao, Lidong Bing, Wang Chen, Michael R Lyu, and Irwin King. 2018. Difficulty controllable generation of reading comprehension questions. *arXiv preprint arXiv:1807.03586*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Sebastian Gombert, Lukas Menzel, 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 (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

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 (HEALTH)*, 3(1):1–23.

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.

Perry N. Halkitis et al. 1996. Estimating Testing Time: The Effects of Item Characteristics on Response Latency. *ERIC*.

Jun He, Li Peng, Bo Sun, Lejun Yu, and Yinghui Zhang. 2021. Automatically predict question difficulty for reading comprehension exercises. In *2021 ieee 33rd international conference on tools with artificial intelligence (ictai)*, pages 1398–1402. IEEE.

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. *Inf. Process. Manag.*, 54:969–984.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *AAAI*, pages 1352–1359.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ghader Kurdi, Bijan Parsia, and Uli Sattler. 2016. An experimental evaluation of automatically generated multiple choice questions from ontologies. In *OWL: Experiences And directions–reasoner evaluation*, pages 24–39. Springer.

Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ekaterina Loginova, Luca Benedetto, Dries Benoit, and Paolo Cremonesi. 2021. Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. In *RANLP 2021*, pages 846–855. INCOMA.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. 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*, pages 3245–3253.

Arya D McCarthy, Kevin P Yancey, Geoffrey T LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 883–899.

Robert J Mislevy. 1988. Exploiting auxiliary information about items in the estimation of rasch item difficulty parameters. *Applied Psychological Measurement*, 12(3):281–296.

Cynthia G. Parshall et al. 1994. Response Latency: An Investigation into Determinants of Item-Level Timing. *ERIC*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

JW Pennebaker, RL Boyd, RJ Booth, A Ashokkumar, and ME Francis. 2022. Linguistic inquiry and word count: Liwc-22. pennebaker conglomerates.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gummuluri Venkata Ravi Ram, Kesanam Ashinee, and Anand Kumar M. 2024. Leveraging Physical and Semantic Features of text item for Difficulty and Response Time Prediction of USMLE 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.

Alvaro Rodrigo, Sergio Moreno-Álvarez, and Anselmo Peñas. 2024. 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)*, Mexico City, Mexico. Association for Computational Linguistics.

Ana-Cristina Rogoz and Radu Tudor Ionescu. 2024. UnibucLLM: Harnessing LLMs for Automated Prediction of Item Difficulty and Item Response Time. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Schneiderand, Haomiao Jin, Bart Orriens, Doerte U. Junghaenel, Arie Kapteyn, Erik Meijer, and Arthur A. Stone. 2023. Using Attributes of Survey Items to Predict Response Times May Benefit Survey Research. *Field Methods*, 35:87–99.

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

Samira Shaikh, Thiago Ferreira, and Amanda Stent, editors. 2022. *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting.

Russell Winsor Smith. 2000. *An exploratory analysis of item parameters and characteristics that influence item level response time*. The University of Nebraska-Lincoln.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Lisa Ann Keller Stowe. 2002. *Small-sample item parameter estimation in the three parameter logistic model: Using collateral information*. Ph.D. thesis, University of Massachusetts Amherst.

Yunik 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.

Hariharan Swaminathan, Ronald K Hambleton, Stephen G Sireci, Dehui Xing, and Saba M Rizavi. 2003. Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied psychological measurement*, 27(1):27–51.

David B. Swanson, Susan M. Case, Douglas R. Ripkey, Brian E. Clauser, and Matthew C. Holtman. 2001. Relationships Among Item Characteristics, Examine Characteristics, and Response Times on USMLE Step 1. *Academic Medicine*, 76:114–116.

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)*, Mexico City, Mexico. Association for Computational Linguistics.

Hariram Veeramani, Natarajan Balaji Shankar Balaji, and Surendrabikram Thapa. 2024. Large Language Model-based Framework for Item Difficulty and Response Time Estimation for Assessments. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

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

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, Peter Baldwin, Christopher Runyon, et al. 2023. Extracting linguistic signal from item text and its application to modeling item characteristics. In *Advancing Natural Language Processing in Educational Assessment*, pages 167–182. Routledge.

Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 2020. Predicting item survival for multiple choice questions in a high-stakesmedical exam. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , Marseille, 11–16 May 2020*, page 6814-6820.

Victoria Yaneva, Daniel Jurich, Peter Baldwin, et al. 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.

Mehrdad Yousefpoori-Naeim, Shayan 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 (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

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.