Question Difficulty Prediction Based on Virtual Test-Takers and Item Response Theory

Masaki Uto[1,*], Yuto Tomikawa[1] and Ayaka Suzuki[1]

[1] The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan

Abstract

Predicting the difficulty of test questions is a crucial task in the field of education. Many recent studies have proposed supervised machine learning methods that predict difficulty from question text. However, this approach requires a large dataset of questions with known difficulties to train difficulty prediction models. Recently, another approach was proposed that uses question-answering (QA) systems as virtual test-takers. This method predicts question difficulty based on the correct/incorrect responses obtained from QA systems, obviating the need to pre-collect questions with known difficulties. However, this approach is limited by the fact that the scale of difficulty values estimated from the responses of QA systems do not necessarily align with the scale derived from human test-takers' responses. To overcome this limitation, we propose a novel method that utilizes QA systems to predict question difficulty while ensuring the difficulty scale aligns with that derived from human test-takers. Our method uses the principle of test linking from item response theory to transform the difficulty scale predicted by QA systems into one derived from human test-takers. Experiments using real data demonstrate that our proposed method can achieve higher accuracy in difficulty prediction compared with conventional methods.

Keywords

Question difficulty, item response theory, large language models, question answering, educational measurement

1. Introduction

Estimating the difficulty of test questions is a crucial task in the education domain. For example, in the context of learning support, providing questions of appropriate difficulty to individual learners enhances learning. Accordingly, such adaptive question presentation is a common objective of intelligent tutoring systems, adaptive learning systems, and knowledge tracing technologies [1, 2, 3, 4, 5, 6, 7]. Furthermore, in the context of educational measurement, estimating question difficulty, specifically using item response theory (IRT) [8], enables sophisticated testing operations, including (1) *adaptive testing*, which enables accurate measurement of ability in a short time by presenting questions with a difficulty tailored to each test-taker's ability [9]; (2) *uniform test assembly*, which involves composing multiple test forms with equivalent difficulty levels [10]; and (3) *test linking*, which facilitates ability estimation on a common scale for

EvalLAC'24: Workshop on Automatic Evaluation of Learning and Assessment Content, July 08, 2024, Recife, Brazil

[*] Corresponding author.

uto@ai.lab.uec.ac.jp (M. Uto)
https://sites.google.com/site/utomasakieng/ (M. Uto)
0000-0002-9330-5158 (M. Uto)

test-takers who have taken different test forms [11, 12]. Given these, it is evident that estimating question difficulty plays a crucial role in various essential tasks in the educational field.

The most common approach for estimating question difficulty entails presenting target questions to human test-takers and using the resulting correct/incorrect response data to estimate their difficulties [11, 12, 13]. Methods for quantifying difficulty are generally divided into two approaches: one based on classical test theory [14], which quantifies question difficulty through the correct answer rate, and another based on IRT [15]. However, regardless of the difficulty quantification approach, this approach necessitates prior administration of target questions to human test-takers, thereby incurring significant costs and potentially compromising the reliability of the test owing to exposure of its content.

Methodologies employing natural language processing technology to predict question difficulty from question texts have recently attracted widespread attention as a means of overcoming this limitation [15, 16, 17, 18, 19, 20, 21]. In this approach, a large dataset of questions with known difficulties is assumed to be given. This dataset is compiled by presenting a large number of questions to a specific group of test-takers and estimating their difficulties from the correct/incorrect responses obtained. The resulting dataset, containing questions with known difficulties, is then used to train a supervised machine learning model that is capable of predicting the difficulties of questions from their texts.

Existing methods based on this approach can be broadly divided into feature-based and neural-based methods [15, 21]. Typical feature-based methods include R2DE (Regressor for Difficulty and Discrimination Estimation) [18] and its extension models (e.g., [16]). However, these methods require meticulous feature engineering to achieve high accuracy. Neural-based methods obviate the necessity for feature engineering by utilizing deep neural networks that process the sequence of words in a question. Recent studies have proposed neural-based methods that utilize pre-trained transformer models such as BERT (bidirectional encoder representations from transformers) [22] and DistilBERT (Distilled-BERT) [23], as illustrated in Fig. 1 [1] [17, 19, 20].

However, even with these neural methods, the accuracy of difficulty prediction often remains modest. For example, a recent study utilizing difficulty prediction models based on BERT and DistilBERT reported that the correlation between the predicted and actual IRT-based difficulty values was not sufficiently high, at 0.441, despite employing relatively large training instances with around 6,700 samples [24]. This suggests that there are inherent limitations to the accuracy of difficulty prediction using this approach, potentially due to the substantial differences between the tasks of question difficulty prediction and general natural-language understanding. These differences complicate the process of transferring the language understanding capability of pre-trained models to question difficulty–prediction tasks.

On the other hand, an alternative approach has been explored that predicts difficulty using question-answering (QA) systems as virtual test-takers [25, 26, 27], as outlined in Fig. 2. This approach constructs several QA systems in advance and predicts the difficulties of target questions

[1] The input for these models consists of the sequence of words in a question text, including related information, such as the reading passage and the correct or distractor options. In the figure, w_t represents the t-th word of an input text sequence, and n denotes the length of the input. The [CLS] symbol signifies a special token, whose output vector serves as a distributed representation of the given text. Consequently, the model predicts a difficulty value by converting this distributed representation vector from BERT or DistilBERT into a scalar value through a linear layer.

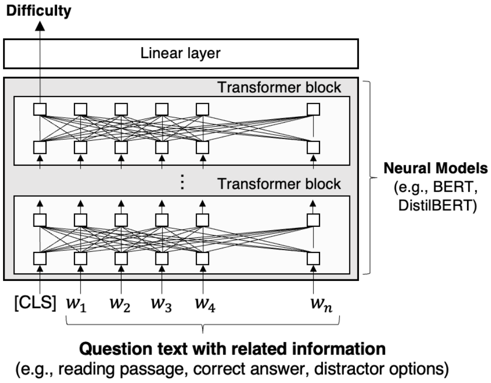

Difficulty
Linear layer
Transformer block
Transformer block
Neural Models
(e.g., BERT,
DistilBERT)
[CLS] w_1 w_2 w_3 w_4 w_n
Question text with related information
(e.g., reading passage, correct answer, distractor options)

Figure 1: Outline of a neural difficulty prediction model.

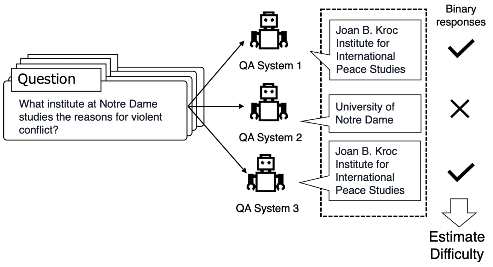

Question
What institute at Notre Dame studies the reasons for violent conflict?
QA System 1
QA System 2
QA System 3
Joan B. Kroc Institute for International Peace Studies
University of Notre Dame
Joan B. Kroc Institute for International Peace Studies
Binary responses
✓
×
✓
Estimate Difficulty

Figure 2: Outline of difficulty prediction using QA systems.

by estimating them from the correct/incorrect responses of the QA systems. For instance, Gao et al. [25] introduced a binary classification method for question difficulty, categorizing questions as "easy" if answered correctly by two QA systems, and "hard" if answered incorrectly by the QA systems. Additionally, Byrd et al. [26] and Uto et al. [27] proposed estimating IRT-based question difficulty by using the correct/incorrect responses from various QA systems. The key premise of this approach is that natural language understanding and QA capabilities are closely related, which makes the construction of QA systems based on pre-trained neural models easier than that of text-based difficulty prediction models. Although this approach can provide a difficulty prediction system without the need to pre-collect questions with known difficulties, it is limited by the fact that the scale of difficulty values estimated from the responses of QA systems will necessarily not align with the scale derived from human test-takers' responses. If these scales do not align, the difficulty values derived from QA systems may not be applicable or meaningful for human test-takers.

To overcome this limitation, we propose a novel method for predicting question difficulty using QA systems while ensuring the difficulty scale aligns with that derived from human test-takers. This method leverages the principle of test linking within IRT [28, 29], a well-designed strategy for unifying scales of IRT parameters estimated from different datasets. Specifically, our method initially collects correct/incorrect responses to a set of questions from human test-takers and various QA systems. IRT is then applied to this response data to estimate their ability values. In this process, we first estimate the ability values of human test-takers using only their response data. Subsequently, given the estimated human test-takers' ability values, we estimate the ability values of the QA systems with IRT using the entirety of the response data. This enables us to map the QA systems' ability estimates onto the scale of human test-takers. We then estimate the difficulty of new questions based on the correct/incorrect responses from the QA systems, using IRT given the QA systems' ability values. Because the ability scale for QA systems is matched to that of human test-takers, the difficulty values for new questions derived from the QA systems are also aligned with the difficulty scale derived from human test-takers. Experiments with real data confirm that this method outperforms conventional methods of predicting question difficulty from question texts.

To our knowledge, this research is the first to focus on using QA systems to predict question difficulty in alignment with the scale of human test-takers, although a similar attempt has now been investigated [30]. Moreover, within the context of test theory, our proposed method introduces a novel test linking technique based on QA systems. This approach could potentially mark a significant advancement over traditional test theory methodologies, as well as in the field of question difficulty prediction research.

2. Quantification of difficulty using IRT

This study employs IRT to quantify question difficulty because our proposed method leverages the advantages of IRT. IRT uses statistical models, called IRT models, to define the probability of each test-taker's response to a question as a function of both their ability and the question's difficulty. This study uses the Rasch model, the simplest IRT model, which defines the probability that test-taker j will answer question i correctly as

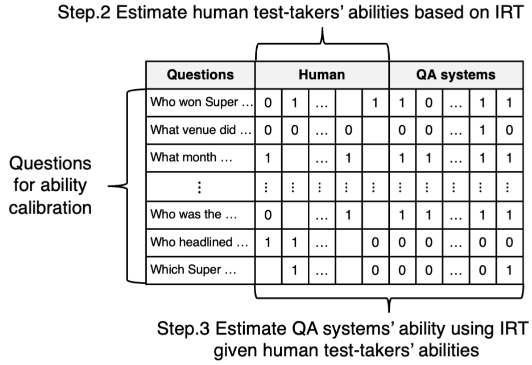
$$p_{ij} = [1 + \exp(-(\theta_j - b_i))]^{-1}, \tag{1}$$

where θ_j is the parameter representing the ability of test-taker j, and b_i is the parameter representing the difficulty of question i. These parameters are estimated from a collection of correct/incorrect responses of a group of test-takers to a set of questions. In the following sections, we assume that difficulty is quantified based on this Rasch model.

3. Proposed method

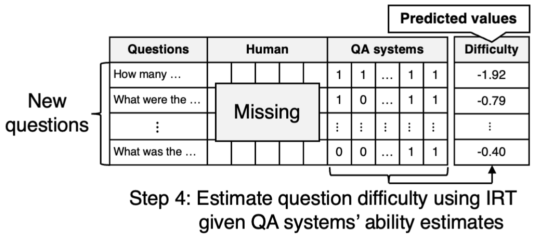
Our proposed method utilizes the concept of test linking based on IRT [28] to realize question difficulty prediction using QA systems while ensuring the predicted difficulty scale aligns with that derived from human test-takers. The detailed processes involve the following steps.

1. Construct QA systems by fine-tuning pre-trained neural models, such as BERT. The fine-tuning process can be performed using publicly available question corpora that match the format of the target questions, or using a small subset of the target question bank. QA systems with varying performances need to be prepared to mimic the diverse abilities of human test-takers, for instance, by varying the base pre-trained models or by limiting the amount of data used for fine-tuning.
2. Administer a set of questions to human test-takers, collect correct/incorrect response data, and apply the Rasch model to the data to estimate the ability values of each test-taker. In the subsequent experiments, we apply expected a posterior (EAP) estimation based on the Markov Chain Monte Carlo algorithm. Note that text-based difficulty prediction methods also require such human response data to construct training data consisting of questions with known difficulty.
3. Gather correct/incorrect responses from QA systems for the same questions administered to human test-takers. Then, using the entirety of the response data collected from both human test-takers and QA systems, estimate the ability values of the QA systems using the Rasch model. In this process, the ability estimates of human test-takers must be given


Step.2 Estimate human test-takers' abilities based on IRT
Questions for ability calibration
Step.3 Estimate QA systems' ability using IRT given human test-takers' abilities

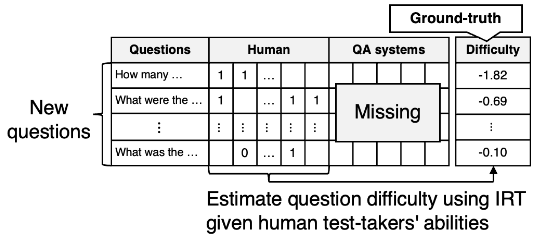
Questions	Human					QA systems				
Who won Super …	0	1	…		1	1	0	…	1	1
What venue did …	0	0	…	0		0	0	…	1	0
What month …	1		…	1		1	1	…	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Who was the …	0		…	1		1	1	…	1	1
Who headlined …	1	1	…		0	0	0	…	0	0
Which Super …		1	…		0	0	0	…	0	1

Figure 3: Preparation phase of our proposed method.


Predicted values
New questions
Missing
Step 4: Estimate question difficulty using IRT given QA systems' ability estimates

Questions	Human					QA systems					Difficulty
How many …						1	1	…	1	1	-1.92
What were the …						1	0	…	1	1	-0.79
⋮						⋮	⋮	⋮	⋮	⋮	⋮
What was the …						0	0	…	1	1	-0.40

Figure 4: Prediction phase of our proposed method.


Ground-truth
New questions
Missing
Estimate question difficulty using IRT given human test-takers' abilities

Questions	Human					QA systems					Difficulty
How many …	1	1	…								-1.82
What were the …	1		…	1	1						-0.69
⋮	⋮	⋮	⋮	⋮	⋮						⋮
What was the …		0	…	1							-0.10

Figure 5: Creation of the ground-truth difficulty.

and fixed. This facilitates the alignment of the QA systems' ability estimates with those of human test-takers on a common scale.

4. Gather correct/incorrect responses from QA systems for new target questions and estimate their difficulty from the response data, using the Rasch model given the QA systems' ability estimates. Because the ability scale for QA systems is matched to that of human test-taker abilities, the difficulty values for new questions derived from the QA systems are also aligned with the difficulty scale derived from human test-takers. Note that this difficulty inference is feasible for both scenarios: individual questions one by one, and all questions simultaneously.

Fig. 3 provides conceptual diagrams for steps 2 and 3, which are preparatory phases for difficulty prediction. Additionally, Fig. 4 provides the diagram for step 4, corresponding to the phase of predicting difficulty for new questions.

The advantage of our method compared with conventional text-based difficulty prediction methods is its efficiency in constructing the question difficulty prediction system. As discussed in Section I, text-based difficulty prediction methods require a large number of questions with known difficulty as a training dataset. This means that a vast number of human responses to many questions are required in advance, which incurs extensive costs. The proposed method can significantly reduce the required amount of such response data because the ability parameter of the Rasch model can be estimated from just several dozen questions [31]. Furthermore, QA systems based on pre-trained transformer models are expected to be constructed using a relatively small number of questions, as discussed in Section I. They can therefore contribute greatly to reducing the necessary data for constructing a question difficulty prediction system.

Furthermore, compared with conventional QA-based difficulty prediction methods, our method has the advantage of being able to predict question difficulty on a scale aligned with that derived from human test-takers' responses, using only the QA systems.

4. Evaluation experiment

We designed an empirical experiment to evaluate the effectiveness of our method. In it, the accuracy of predicting difficulty for new questions using our method is compared with that of conventional text-based difficulty prediction methods.

4.1. Data

For our experiments, we utilized two publicly available datasets: EVKD (The ESL Learners' Vocabulary Knowledge Dataset) [32, 33], which comprises English vocabulary tests, and SQuAD (The Stanford Question Answering Dataset) [34], commonly employed as a benchmark in QA and question generation research.

EVKD comprises English vocabulary test questions that require test-takers to select the appropriate expression to replace a specified part of a given English sentence from multiple choices. The dataset contains 100 questions, each with question text, one correct answer choice, and three distractor choices, along with correct/incorrect response data from 100 English learners. In this experiment, a randomly selected subset of 50 questions was used to construct QA systems, while the remaining 50 questions and their corresponding response data were used to evaluate the prediction performance of both our proposed and conventional methods.

SQuAD is a dataset for reading comprehension, comprising reading passages, comprehension questions, and reference answers. The reading passages are sourced from Wikipedia, with questions and reference answers generated by crowdworkers. Each reference answer corresponds to a segment of the text in the reading passage. The SQuAD dataset is pre-split into 90% for training and 10% for testing. However, it cannot be directly applied to our experiment as it lacks correct/incorrect response data from human test-takers for the questions it contains. Thus, for this study, we randomly selected 570 questions from the SQuAD test dataset and collected response data from 10 human test-takers for these questions. On average each test-taker answered 120 questions, guaranteeing that at least two test-takers responded to each question. Answer correctness was verified by exact match after preprocessing the test-takers' answers (e.g., removing articles, standardizing case, eliminating spaces). In this experiment, the SQuAD training dataset was used to construct QA systems, and the 570 questions with responses from human test-takers were used to evaluate the prediction performance of both our proposed and conventional methods.

4.2. QA systems

For each dataset, a variety of QA systems with differing abilities were developed. Specifically, we utilized 12 pre-trained transformer models from Huggingface[2]: bert-base-uncased, bert-large-uncased, roberta-base, roberta-large, microsoft/deberta-base, microsoft/deberta-large, microsoft/deberta-v3-base, microsoft/deberta-v3-large, albert-base-v1, albert-base-v2, albert-large-v2, and distilbert-base-uncased. We adapted the output layers of these models to align with the question type of each dataset, and conducted model training with varying amounts of

[2] https://huggingface.co/

data to generate QA systems of diverse performance levels[3].

For the EVKD dataset, the QA systems were designed as classifiers that process the question text and four choice options to identify the correct answer. Specifically, the special token [CLS] is appended to the input text, and an output classification layer is added atop the output vector corresponding to this token. In addition to appending the [CLS] token at the beginning of the input, a special token [SEP] is inserted as a separator between the question text and the four choice options. As mentioned previously, the data from 50 questions was available for constructing the QA systems. Accordingly, we trained each QA system using the entire dataset and random subsets corresponding to 40, 30, 20, 10, and 5 questions, respectively.

For the SQuAD dataset, the QA systems were configured to predict the start and end positions of the answer within the reading passage. The input for the models comprises the concatenation of a passage and a question text, separated by the special token [SEP]. We trained each QA system using the entire SQuAD training dataset and random subsets of 3000, 2400, 1800, 1200, and 600 data points, respectively.

This procedure resulted in 72 QA systems with varying levels of ability for each dataset.

4.3. Experimental procedure

We evaluated the performance of difficulty prediction using both our proposed method with constructed QA systems and conventional methods that predict difficulty from question texts using supervised regression models. As detailed in Section 4.1, for performance evaluation, we can utilize 50 questions from the EVKD dataset and 570 questions from the SQuAD dataset, along with the respective responses from human test-takers. Thus, for each dataset we randomly split the data into 90% and 10%. The 90% portion, denoted as $\boldsymbol{D}$, was used to develop difficulty predictors. Developing difficulty predictors corresponds to the process of estimating the abilities of human test-takers and QA systems in our method as well as that of training a regression model for difficulty prediction from question texts in the conventional method. The remaining 10%, denoted as $\boldsymbol{E}$, was used to evaluate the accuracy of the difficulty prediction.

Specifically, in our method, the ability values of test-takers were initially estimated using the Rasch model based on the correct/incorrect response data from human test-takers within $\boldsymbol{D}$. Subsequently, the ability values of the QA systems were estimated using response data from both the 72 QA systems and human test-takers for the same questions, while fixing the ability values of human test-takers. Finally, for each question in $\boldsymbol{E}$, the difficulty was estimated using the Rasch model based on responses from the QA systems, with the ability estimates of the QA systems held fixed. These calculated values were considered as the predicted difficulty values.

In the conventional method, the difficulty of questions within $\boldsymbol{D}$ was first estimated based on the Rasch model using the response data from human test-takers. Subsequently, regression models for predicting difficulty from question texts were trained using the set of questions with estimated difficulties[4]. We explored two neural regression models, BERT and DistilBERT, which were also utilized in prior research. For each question in $\boldsymbol{E}$, the predicted difficulty values were derived by inputting the question texts into these trained models.

[3] The training of the QA systems employed AdamW with a learning rate of 1e-5 and a maximum of 5 epochs. Neither the EVKD nor the SQuAD datasets were used in the original pretraining of each transformer model.

[4] The training was done by AdamW with a learning rate of 1e-5 and a maximum of 10 epochs.

Table 1
Experimental results

		Correlation coefficient			Regression coefficient		
		Prop.	BERT	DistilBERT	Prop.	BERT	DistilBERT
EVKD	Mean	**0.326**	0.274	0.275	**0.215**	0.034	0.026
	SD	0.251	0.233	0.412	0.175	0.030	0.031
SQuAD	Mean	**0.588**	0.191	0.134	**1.339**	0.065	0.034
	SD	0.054	0.043	0.047	0.125	0.019	0.014

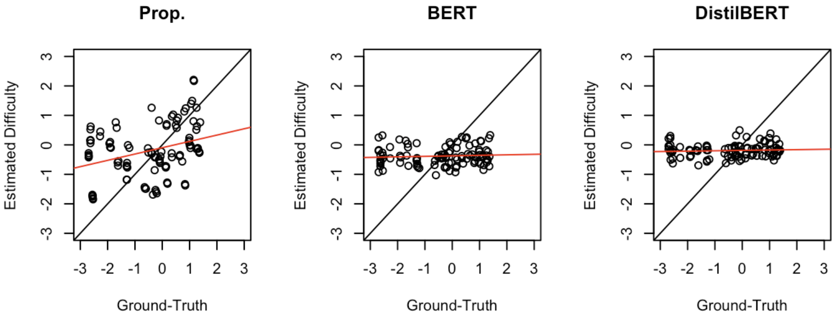


Figure 6: Relationship between predicted difficulty values and ground-truth in the EVKD dataset.

Because the objective of this study was to predict difficulty values for new questions that align with human scales, the ground-truth difficulty values for each question in $\boldsymbol{E}$ were estimated from the response data of human test-takers within $\boldsymbol{E}$. In this difficulty estimation process, the ability values of human test-takers, estimated from $\boldsymbol{D}$, were given. The process for generating the ground-truth difficulty values is depicted in Fig. 5.

We evaluated the accuracy of difficulty prediction by comparing the predicted difficulty for each question in $\boldsymbol{E}$ provided by each method to the corresponding ground-truth defined above. Correlation coefficients and regression coefficients served as metrics for evaluating prediction accuracy. A higher correlation approaching one and regression coefficient values nearing one signify enhanced prediction accuracy. To improve the reliability of the experimental results, we repeated the experiment 10 times, varying the random splits of $\boldsymbol{D}$ and $\boldsymbol{E}$ each time.

4.4. Experimental results

The experimental results are presented in Table 1. The rows labeled mean and SD represent the average performance over 10 repetitions and its standard deviation, respectively. The results demonstrate that our proposed method outperforms conventional methods in both datasets. Notably, when examining the regression coefficients, we can see that the values for conventional methods are nearly zero. To elucidate this phenomenon, Fig. 6 and Fig. 7 display scatter plots of the predicted difficulty values against the ground truth for each dataset. These figures demonstrate how conventional methods produce limited variances in predicted difficulties, failing to accurately capture the range of difficulty.

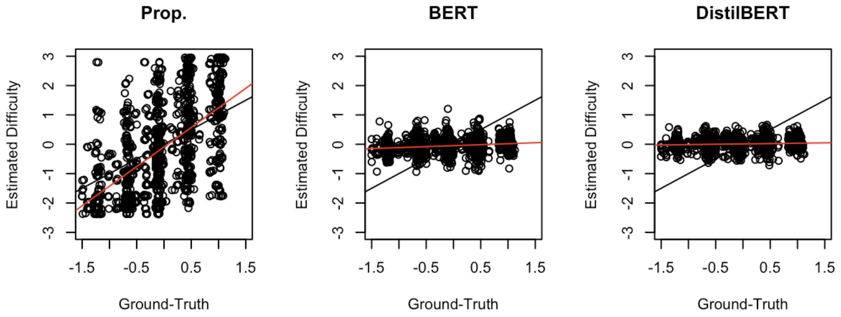


Figure 7: Relationship between predicted difficulty values and ground-truth in the SQuAD dataset.

Table 2
Ability estimates and their PSD for human test-takers and for each of the 12 pre-trained transformer models used in QA systems.

	Ability estimates θ_j		PSD of θ_j	
Test-takers	Mean	SD	Mean	SD
human	0.056	0.588	0.243	0.034
distilbert.base.uncased	-0.189	1.602	0.133	0.017
bert.base.uncased	0.374	1.161	0.130	0.011
albert.base.v1	1.025	0.733	0.134	0.013
albert.large.v2	1.031	0.510	0.133	0.006
albert.base.v2	1.101	0.766	0.138	0.012
bert.large.uncased	1.186	0.803	0.138	0.014
roberta.base	1.375	0.814	0.140	0.015
deberta.base	1.542	0.684	0.144	0.014
deberta.v3.base	1.590	0.646	0.143	0.014
roberta.large	1.789	0.587	0.148	0.013
deberta.v3.large	1.993	0.579	0.150	0.014
deberta.large	2.028	0.384	0.152	0.010

Table 3
Ability estimates of QA systems across various training sample sizes.

Sample	Ability estimates θ_j	
size	Mean	SD
300	0.136	1.197
600	0.816	0.747
900	1.194	0.609
1200	1.362	0.610
1500	1.516	0.563
full	2.398	0.403

4.5. Additional analysis

This section analyzes the ability estimates of the QA systems constructed for our method. Specifically, we investigated the estimated abilities of the 72 QA systems and the human test-takers, which were obtained from the experiments conducted using the SQuAD dataset. Because the experiments yielded ability estimates for each of the 10 repetitions, we first confirmed the correlations and root mean squared errors (RMSEs) in the estimates among all pairs of the 10 repetitions. We found that the average correlation was 0.995 with an SD of 0.001, while the average RMSE was 0.102 with an SD of 0.014. These results suggest that the ability estimates are strongly consistent among the repetitions.

Thus, we subsequently investigated the ability estimates obtained from the first repetitions. Table 2 shows the statistics corresponding to the ability estimates for human test-takers and for each of the 12 pre-trained transformer models used in the 72 QA systems. The statistics include the average and SD of the ability estimates as well as those of the posterior standard deviations

(PSDs) for the ability estimates. This table reveals some reasonable trends. Specifically, variants of the DeBERTa model, one of the latest models, exhibit higher average abilities, while the distilBERT, a simplified version of BERT, shows the lowest average abilities. Furthermore, when comparing the sizes of each model, the larger models tend to provide higher abilities. Table 2 also indicates that the PSDs are low for all test-takers, including humans and QA systems, suggesting that the accuracy of ability estimation would be acceptable.

Furthermore, Table 3 shows the average and SD of the ability estimates of the QA systems across the various training sample sizes, demonstrating that an increase in training sample size leads to an increase in ability estimates, which is also a reasonable trend.

Finally, the analysis of Tables 2 and 3 indicates that the developed QA systems tend to have higher abilities than the human test-takers, suggesting a mismatch in the ability distribution between QA systems and human test-takers. This discrepancy can potentially lead to a deterioration in difficulty estimation, especially over the small difficulty value range. Therefore, filtering the QA systems or adding relatively weak QA systems could be beneficial for improving the performance of difficulty estimation, a task we intend to focus on in future studies.

5. Conclusion

In this study, we introduced an IRT-based question difficulty prediction method that uses QA systems as virtual test-takers, while ensuring alignment of the difficulty scale with that derived from human test-takers. Through experiments with real data, we showed that our proposed method outperforms traditional text-based difficulty prediction methods.

This study has some limitations. The first is the scarcity of detailed experiments, which is due primarily to the limited availability of open datasets that include both question data and human test-takers' responses. Future work that evaluates the effectiveness of the proposed method across a broader range of datasets in various educational domains to identify necessary adaptations. Furthermore, possible further investigations based on our data, such as examining if larger difficulty estimation errors are typically made against lower-difficulty questions, will also be part of our future research.

Second, the proposed method necessitates collecting questions for training QA systems in addition to those that include human test-takers' responses. We assume that it is generally easier to collect questions without human responses than those with, and a relatively small dataset may suffice for training QA systems. However, the feasibility of this data collection process and the amount of data required for training should be examined in future investigations. Furthermore, a recent study proposes a method that considers the uncertainty of predictions from a QA system as the difficulty for multiple-choice questions [35]. This idea might be integrated with our approach to enhance its effectiveness.

Finally, it is anticipated that the proposed method will require significantly fewer questions with human responses compared with the conventional text-based difficulty prediction approach. This is because the proposed method uses the response data primarily to estimate a small number of parameters in an IRT model, whereas the conventional approach uses these data to train a large neural model on a complex task. Future work will explore the extent to which the proposed method can reduce the amount of data required and the corresponding costs.

References

[1] G. Kurdi, J. Leo, B. Parsia, U. Sattler, S. Al-Emari, A systematic review of automatic question generation for educational purposes, International Journal of Artificial Intelligence in Education 30 (2019) 121–204.

[2] N.-T. Le, T. Kojiri, N. Pinkwart, Automatic question generation for educational applications – the state of art, Advanced Computational Methods for Knowledge Engineering 282 (2014) 325–338.

[3] M. Rathod, T. Tu, K. Stasaski, Educational multi-question generation for reading comprehension, in: Proc. Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, Washington, 2022, pp. 216–223.

[4] R. Zhang, J. Guo, L. Chen, Y. Fan, X. Cheng, A review on question generation from natural language text, ACM Transactions on Information Systems 40 (2021) 1–43.

[5] M. Liu, R. A. Calvo, Using information extraction to generate trigger questions for academic writing support, in: Intelligent Tutoring Systems, Chania, Crete, Greece, 2012, pp. 358–367.

[6] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, J. Sohl-Dickstein, Deep knowledge tracing, Advances in neural information processing systems 28 (2015) 505–513.

[7] Z. Wang, X. Feng, J. Tang, G. Y. Huang, Z. Liu, Deep knowledge tracing with side information, in: Proc. International conference on artificial intelligence in education, 2019, pp. 303–308.

[8] F. M. Lord, Applications of item response theory to practical testing problems, Routledge, 1980.

[9] W. J. van der Linden, P. J. Pashley, Computerized Adaptive Testing: Theory and Practice, Springer Netherlands, 2000.

[10] D. Belov, Uniform test assembly: Concepts, problems, solvers, and applications for adaptive testing, Journal of Computerized Adaptive Testing 5 (2017) 1–21.

[11] F. B. Baker, S. H. Kim, Item Response Theory: Parameter Estimation Techniques, CRC Press, Boca Raton, FL, USA, 2004.

[12] W. J. van der Linden, R. K. Hambleton, Handbook of modern item response theory, Springer Verlag, 1996.

[13] M. J. Kolen, R. L. Brennan, Test Equating, Scaling, and Linking: Methods and Practices, Springer New York, 2013.

[14] F. M. Lord, M. R. Novick, Statistical theories of mental test scores, Information Age Publishing, 1968.

[15] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, ACM Computing Surveys 55 (2023).

[16] L. Benedetto, G. Aradelli, P. Cremonesi, A. Cappelli, A. Giussani, R. Turrin, On the application of transformers for estimating the difficulty of multiple-choice questions from text, in: Proc. Workshop on Innovative Use of NLP for Building Educational Applications, 2021, pp. 147–157.

[17] L. Benedetto, A. Cappelli, R. Turrin, P. Cremonesi, Introducing a framework to assess newly created questions with natural language processing, in: Proc. International Conference on Artificial Intelligence in Education, 2020, pp. 43–54.

[18] L. Benedetto, A. Cappelli, R. Turrin, P. Cremonesi, R2DE: a NLP approach to estimating IRT parameters of newly generated questions, in: Proc. International Conference on Learning Analytics & Knowledge, 2020, pp. 412–421.

[19] A. D. McCarthy, K. P. Yancey, G. T. LaFlair, J. Egbert, M. Liao, B. Settles, Jump-starting item parameters for adaptive language tests, in: Proc. Conference on Empirical Methods in Natural Language Processing, 2021, pp. 883–899.

[20] K. Xue, V. Yaneva, C. Runyon, P. Baldwin, Predicting the difficulty and response time of multiple choice questions using transfer learning, in: Proc. Workshop on Innovative Use of NLP for Building Educational Applications, 2020, pp. 193–197.

[21] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, Text-based question difficulty prediction: A systematic review of automatic approaches, International Journal of Artificial Intelligence in Education (2023).

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[23] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv (19).

[24] L. Benedetto, A quantitative study of nlp approaches to question difficulty estimation, in: Proc. International Conference on Artificial Intelligence in Education, 2023, pp. 428–434.

[25] Y. Gao, L. Bing, W. Chen, M. Lyu, I. King, Difficulty controllable generation of reading comprehension questions, in: Proc. International Joint Conference on Artificial Intelligence, 2019, pp. 4968–4974.

[26] M. Byrd, S. Srivastava, Predicting difficulty and discrimination of natural language questions, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 119–130.

[27] M. Uto, Y. Tomikawa, A. Suzuki, Difficulty-controllable neural question generation for reading comprehension using item response theory, in: Proc. Workshop on Innovative Use of NLP for Building Educational Applications, 2023, pp. 119–129.

[28] M. J. Kolen, R. L. Brennan, Test Equating, Scaling, and Linking, Springer Verlag, 2014.

[29] M. Uto, Accuracy of performance-test linking based on a many-facet Rasch model, Behavior Research Methods 53 (2021) 1440–1454.

[30] H. Maeda, Field-testing multiple-choice questions with AI examinees, Preprint available at Research Square, 2024.

[31] J. M. Linacre, Sample size and item calibration stability, Rasch measurement transactions 7 (1994).

[32] Y. Ehara, Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing, in: Proc. Language Resources and Evaluation Conference, 2018, pp. 484–488.

[33] Y. Ehara, I. Sato, H. Oiwa, H. Nakagawa, Mining words in the minds of second language learners: Learner-specific word difficulty, in: Proc. International Conference on Computational Linguistics, 2012, pp. 799–814.

[34] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proc. Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.

[35] E. Loginova, L. Benedetto, D. Benoit, P. Cremonesi, Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2021, pp. 846–855.