Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

# Difficulty-Controllable Reading Comprehension Question Generation Considering the Difficulty of Reading Passages

**Yuto TOMIKAWA*** **& Masaki UTO***
*The University of Electro-Communications, Japan*
*{tomikawa, uto}@ai.lab.uec.ac.jp

**Abstract:** In recent years, various question generation (QG) methods for reading comprehension have been proposed that automatically generate questions related to given reading passages. Specifically, QG methods based on deep neural networks have succeeded in generating high-quality questions. To apply such QG methods in educational systems, such as intelligent tutoring systems and adaptive learning systems, it is crucial to generate questions with difficulty levels that are appropriate for each learner's reading ability. To meet this need, several difficulty-controllable QG methods have been proposed recently. However, a limitation of existing difficulty-controllable methods is that they overlook the difficulty of the reading passages, which are given as the input context for QG. Since the difficulty of reading passages can affect the difficulty of the generated questions, selecting reading passages with appropriate difficulty is crucial. Therefore, in this study, we develop a difficulty-controllable QG method that includes a mechanism for selecting reading passages with appropriate difficulty for each learner. Our approach begins with the proposal of a new item response theory (IRT) model capable of simultaneously estimating the difficulty of both questions and reading passages. Using the developed IRT model and the latest IRT-based difficulty-controllable QG method, we propose a framework to select reading passages and generate questions that are appropriate for each learner's reading ability.

**Keywords:** Question generation, reading comprehension, item response theory, deep neural networks, natural language process, large language models

## 1. Introduction

In recent years, with the emergence of generative AI, the development of reading comprehension abilities has become increasingly important in education. To cultivate these abilities, it is effective to ask learners to read various reading materials while answering corresponding comprehension questions (Kurdi et al., 2020; Le et al., 2014; Rathod et al., 2022; Zhang et al., 2021). This approach helps focus learners' attention on the content and provides an opportunity to identify any misconceptions they might have (Kurdi et al., 2020), thereby supporting the development of reading comprehension abilities. However, manually creating a variety of comprehension questions for different reading materials is both time-consuming and costly. To address this issue, automatic question generation (QG) techniques have gained attention in recent years (Mostow & Chen, 2009; Heilman & Smith, 2010; Tomikawa & Uto, 2024). Specifically, QG methods based on deep neural networks have recently succeeded in generating high-quality questions (Du et al., 2017; Perkoff et al., 2023)

To effectively apply such QG methods in educational systems, such as intelligent tutoring systems (Graesser et al., 2012) and adaptive learning systems, it is more beneficial to generate questions with difficulty levels tailored to each learner's reading ability rather than generating them randomly. To achieve this, recent studies have proposed techniques for generating reading comprehension questions with controllable difficulty levels (Cheng et al.,

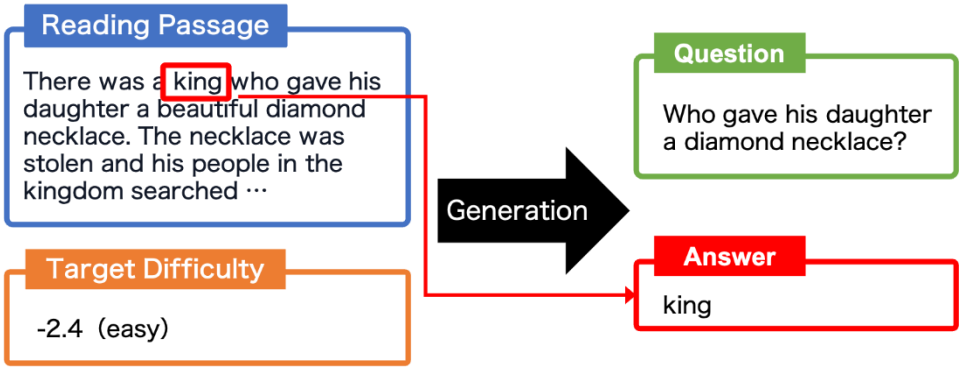


*Figure 1.* Outline of Difficulty Controllable Question Generation

2021; Gao et al., 2019; Uto et al., 2023; Goto et al., 2024; Tomikawa & Uto, 2024). For example, Uto et al. (2023) proposed a difficulty-controllable neural QG method that takes the difficulty level of questions, quantified using item response theory (IRT) (Lord, 1980), along with the reading passages as input, and outputs corresponding questions. Furthermore, Tomikawa and Uto (2024) extended this approach to a multiple-choice QG method. However, those existing methods focus only on controlling the difficulty of the questions, while overlooking the fact that the difficulty of the reading passages can affect the difficulty of the generated questions. For example, it is likely more challenging to generate high-difficulty questions from an easy reading passage compared to generating high-difficulty questions from a difficult passage. Additionally, selecting reading passages that are appropriate for each learner is crucial for enhancing their motivation and engagement.

Therefore, in this study, we develop a difficulty-controllable QG method that includes a mechanism for selecting reading passages with appropriate difficulty for each learner. Our approach begins with the proposal of a new IRT model that extends the Rasch model, one of the traditional IRT models, into a hierarchical Bayesian model to estimate the difficulty of reading passages alongside the difficulty of questions. Using the newly developed IRT model and the conventional IRT-based difficulty-controllable QG method proposed by Uto et al. (2023), we propose a framework to select reading passages and generate questions that are appropriate for each learner's reading ability.

In this study, we conducted experiments using a benchmark dataset and demonstrated that the difficulty of the reading passages influences the difficulty controllability of QG. Our experiments further show that selecting reading passages and generating questions with difficulty levels appropriate for each learner's ability is effective for efficiently estimating learner ability and facilitating more accurate difficulty control.

## 2. Conventional Difficulty-Controllable Question Generation Method

This section explains the difficulty-controllable reading comprehension QG method proposed by Uto et al. (2023), which is used as the base method for our study.

The QG task tackled in this study is outlined in Figure 1. As shown in the figure, the task involves generating a reading comprehension question and a corresponding correct answer, given a reading passage and a target difficulty value. Here, we assume that the correct answer to each question consists of a segment of text from the corresponding reading passage, as is typical in answer-aware QG tasks (Du et al., 2017; Zhou et al., 2017; Subramanian, 2018; Sun, 2018).

To develop a QG method for this task, this study assumes the use of the SQuAD dataset (Rajpurkar et al., 2016), which is widely used as a benchmark dataset for question answering and QG tasks (Raffel et al., 2020; Lewis; 2020; He et al., 2021; Goto et al., 2024). The SQuAD dataset consists of approximately 100,000 sets of reading passages, questions created for those passages, and their corresponding answers. The reading passages are collected from Wikipedia, and the questions and answers were generated by crowd workers.

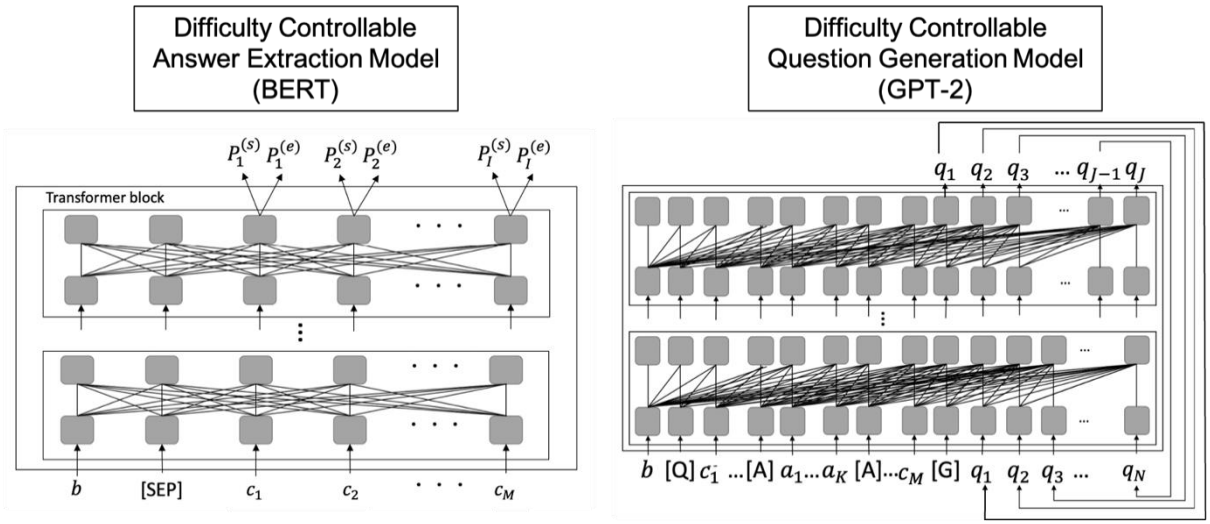


*Figure 2.* Architecture of the Conventional Difficulty-Controllable Question Generation Method

The SQuAD dataset can be represented as $D = \{\boldsymbol{c}_i, \boldsymbol{q}_i, \boldsymbol{a}_i \mid I \in \{1, \dots, I\}\}$. Here, $\boldsymbol{c}_i = \{c_{im} \mid m \in \{1, \dots, M_i\}\}$ denotes the $i$ -th reading passage, $\boldsymbol{q}_i = \{q_{in} \mid n \in \{1, \dots, N_i\}\}$ represents the corresponding question, and $\boldsymbol{a}_i = \{a_{io} \mid o \in \{1, \dots, O_i\}\}$ corresponds to the answer associated with the reading passage and question. Additionally, $c_{im}$, $q_{in}$, and $a_{io}$ represent the $m$-th, $n$-th, and $o$-th words in $\boldsymbol{c}_i$, $\boldsymbol{q}_i$, and $\boldsymbol{a}_i$, respectively, while $M_i$, $N_i$, and $O_i$ indicate the number of words in $\boldsymbol{c}_i$, $\boldsymbol{q}_i$, and $\boldsymbol{a}_i$. $I$ represents the number of data points. In SQuAD, the answer $\boldsymbol{a}_i$ is a substring of the sequence of words in the reading passage $\boldsymbol{c}_i$, i.e., $\boldsymbol{a}_i \subset \boldsymbol{c}_i$.

Although the SQuAD dataset does not include question difficulty levels, constructing difficulty-controllable QG methods requires a dataset of quadruplets $(\boldsymbol{c}_i, \boldsymbol{q}_i, \boldsymbol{a}_i, b_i)$, where $b_i$ represents the difficulty of the $i$-th question. To address this, Uto et al. (2023) began by proposing a method to estimate the difficulty of each question in the SQuAD dataset. Specifically, they proposed to offer each question $\boldsymbol{q}_i$ in the dataset to a number of question answering (QA) systems with varying ability levels, and estimate the difficulty $b_i$ of each question from their correct/incorrect response data. Then, by adding the estimated difficulty values to the dataset, we can get a dataset of quadruplets $(\boldsymbol{c}_i, \boldsymbol{q}_i, \boldsymbol{a}_i, b_i)$. Note that, to quantify the difficulty value, the Rasch model, an IRT model explained in the next section, is used.

Using this extended SQuAD dataset with difficulty levels, the following two models are trained to construct the difficulty-controllable QG.

### *2.1 Difficulty-Controllable Answer Extraction Model*

The first model is the difficulty-controllable answer extraction model, which is designed to extract answers from a reading passage that align with a desired difficulty level. Specifically, this model is implemented as a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2019) that takes as input a concatenated string of the difficulty $b$ and the reading passage $\boldsymbol{c}$, and outputs the start and end positions of the answer in the passage. The architecture of this model is illustrated on the left side of Figure 2. The model fine-tuning is conducted by minimizing the following loss function.

$$Loss^{(ae)} = -\sum_{i=1}^{I}\sum_{m=1}^{M_i}\left\{Z_{im}^{(s)}\log H_{im}^{(s)} + Z_{im}^{(e)}\log H_{im}^{(e)}\right\} \tag{1}$$

$$H_{im}^{(s)} = softmax(\boldsymbol{S}\cdot\boldsymbol{T}_{im}) = \frac{\exp(\boldsymbol{S}\cdot\boldsymbol{T}_{im})}{\sum_{m'=1}^{M_i}\exp(\boldsymbol{S}\cdot\boldsymbol{T}_{im'})} \tag{2}$$

$$H_{im}^{(e)} = softmax(\boldsymbol{E} \cdot \boldsymbol{T}_{im}) = \frac{\exp(\boldsymbol{E} \cdot \boldsymbol{T}_{im})}{\sum_{m'=1}^{M_i} \exp(\boldsymbol{E} \cdot \boldsymbol{T}_{im'})} \quad (3)$$

Here, $Z_{im}^{(s)}$ and $Z_{im}^{(e)}$ are dummy variables that take the value of 1 when the $m$-th word in the $i$-th reading passage is the start position and end position of the answer, respectively. $\boldsymbol{T}_{im}$ denotes the output vector of BERT for the $m$-th word in the reading passage, and $\boldsymbol{S}$ and $\boldsymbol{E}$ represent the learnable weight.

### *2.2 Difficulty-Controllable Question Generation Model*

Another model is the difficulty-controllable QG model, which is designed to generate questions based on a reading passage and an answer that align with a desired difficulty level. Specifically, this model is implemented as a fine-tuned GPT-2 (Generative Pre-trained Transformer 2) model (Radford et al., 2019) that takes as input a concatenated string of the difficulty $b$, the reading passage $\boldsymbol{c}$, and the answer $\boldsymbol{a}$, and generates a question $\boldsymbol{q}$. The architecture of this model is illustrated on the right side of Figure 2. The model fine-tuning is conducted by minimizing the following loss function.

$$Loss^{(qg)} = -\sum_{i=1}^{I}\sum_{n=1}^{N_i} \log\{P(q_{in}|q_{i1},\ \ldots,\ q_{i(n-1)},\ \boldsymbol{c}_i,\ \boldsymbol{a}_i,\ b_i)\} \quad (4)$$

$$P(q_{in}|q_{i1},\ \ldots,\ q_{i(n-1)},\ \boldsymbol{c}_i,\ \boldsymbol{a}_i,\ b_i) = softmax\left(\boldsymbol{G} \cdot \boldsymbol{T}_{pre_{i(n-1)}}^{q_{in}}\right) = \frac{\exp\left(\boldsymbol{G} \cdot \boldsymbol{T}_{pre_{i(n-1)}}^{q_{in}}\right)}{\sum_{v=1}^{V} \boldsymbol{G} \cdot \boldsymbol{T}_{pre_{i(n-1)}}^{q_{iv}}} \quad (5)$$

Here, $V$ is the total vocabulary size of GPT-2, $\boldsymbol{T}_{pre_{i(n-1)}}^{q_{in}}$ is the output vector of GPT-2 corresponding to $\boldsymbol{q}_{in}$ given the word sequence $pre_{i(n-1)} = (q_{i1},\ \ldots,\ q_{i(n-1)},\ \boldsymbol{c}_i,\ \boldsymbol{a}_i,\ b_i)$, and $\boldsymbol{G}$ represents the learnable weight.

## 3. Item Response Theory

As mentioned in Section 2, this study assumes that question difficulty is quantified using IRT. IRT is a test theory that utilizes mathematical models known as IRT models. The IRT models separate the representation of question difficulty and examinee ability, enabling an accurate estimation of examinee ability independent of the specific characteristics of the questions. The Rasch model, the most traditional IRT model, is defined by the following equation.

$$P_{ij}(u_{ij} = 1|\theta_j,\ b_i) = \frac{1}{1 + \exp\left(-(\theta_j - b_i)\right)} \quad (6)$$

Equation (6) represents the probability that a learner $j$ with ability $\theta_j$ will correctly answer a question $i$ with difficulty $b_i$, where, $u_{ij}$ is a variable that takes the value of 1 when learner $j$ answers question $i$ correctly, and takes 0 otherwise. These parameters are estimated from a collection of correct/incorrect response data of learners to questions.

Figure 3 shows the item characteristic curves (ICCs) based on Equation (6) for three questions with different difficulty levels. The horizontal axis represents the ability value $\theta_j$, and the vertical axis represents the probability $P_{ij}$. The three curves correspond to the ICCs for three questions with different difficulty values. From this figure, it can be observed that when $\theta_j$ is the same, the probability of a correct response is higher for questions with a lower difficulty $b_i$.

The difficulty-controllable QG method introduced above uses IRT because it has the unique feature of formulating the relationship between learner ability and question difficulty, making it easier to select difficulty values appropriate to a learner's ability. According to Ueno and Miyazawa (2018), presenting questions at a difficulty level where the learner has a 50% chance of answering correctly is effective for facilitating learning. As shown in Figure 3, the

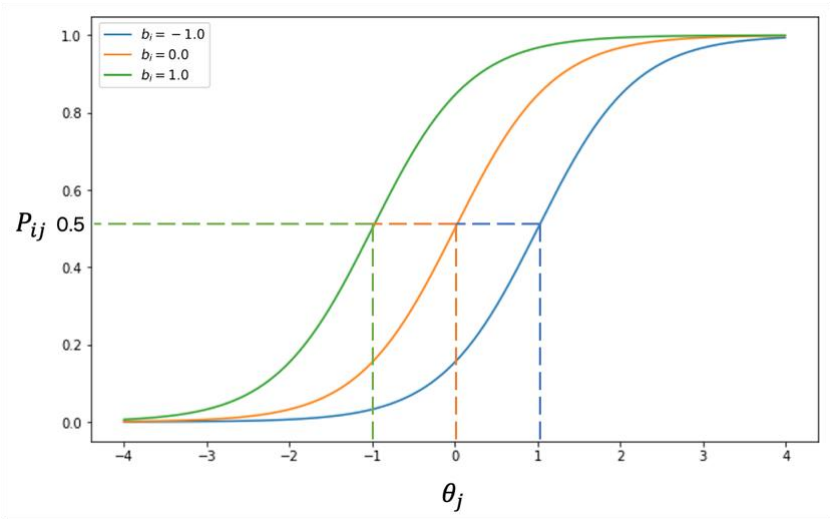


*Figure 3.* Item Characteristic Curves for a Rasch Model with Different Item Difficulty Values

probability of a correct response becomes 0.5 when $b_i = \theta_j$. This suggests that the selecting difficulty $b_i = \theta_j$ is appropriate when generating questions using the difficulty-controllable QG method.

IRT is also utilized as a foundational technology in computerized adaptive testing (CAT) (Van der Linden & Glas, 2010; Ueno & Miyazawa, 2018), a testing method that involves a cycle of sequentially presenting questions at a difficulty level suited to a learner's ability and estimating their ability based on their responses. Specifically, because the standard error of ability estimates is calculable as the inverse square root of the Fisher information (FI), IRT-based CAT generally selects questions repeatedly that have maximum FI for a current ability estimate of each individual learner. In the Rasch model, the FI is maximized when $b = \theta$. Therefore, we can implement the CAT framework through the sequential procedures of estimating ability from the response history and presenting questions with difficulty values close to the current ability estimates. Based on this strategy, Uto et al. (2023) developed an adaptive QG framework that generates questions with difficulty levels appropriate for learners while efficiently estimating their abilities each time a question is presented. As described in Section 5.2, we also employ this adaptive framework to select reading passages and generate questions with difficulty levels appropriate for each learner's ability.

## 4. Proposed Method

The difficulty-controllable QG methods introduced above focus only on the difficulty of the questions, while ignoring the difficulty of the reading passages. However, as discussed in Section 1, selecting reading passages with appropriate difficulty is crucial for enhancing the controllability of question difficulty and the engagement of learners. Therefore, in this study, we develop a difficulty-controllable QG method that includes a mechanism for selecting reading passages with difficulty levels tailored to each learner. For our method, we develop a new IRT model capable of simultaneously estimating the difficulty of both questions and reading passages. Using the proposed IRT model and the IRT-based difficulty-controllable QG method, we propose a framework to select reading passages and generate questions that are appropriate for each learner's reading ability.

### *4.1 Hierarchical IRT Model for Estimating the Difficulty of Reading Passages*

The proposed IRT model, which can estimate the difficulty of both reading passages and questions, is formulated as a hierarchical Bayesian extension of the Rasch model. Our model

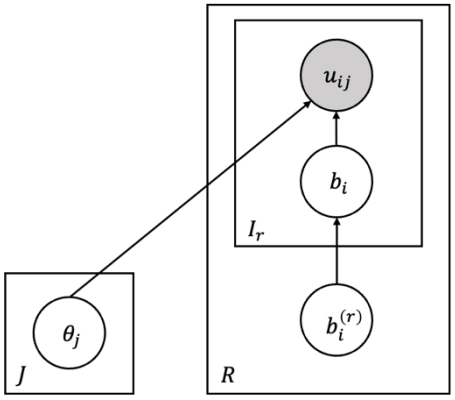


*Figure 4.* The Graphical Model of the Proposed IRT Model.

defines the probability that a learner $j$ with ability $\theta_j$ will correctly answer a question $i$ with difficulty $b_i$, which is related to the reading passage $r$ with difficulty $b_i^{(r)}$, as follows.

$$P_{ij}^{(r)} = \frac{1}{1 + \exp\left(-\left(\theta_j - b_i\right)\right)} \tag{7}$$

$$\theta_j \sim \mathcal{N}(0, 1) \tag{8}$$

$$b_i \sim \mathcal{N}\left(b_i^{(r)}, \sigma\right) \tag{9}$$

$$b_i^{(r)} \sim \mathcal{N}(0, 1) \tag{10}$$

where, $\sigma$ is a hyperparameter, and $\mathcal{N}(m, s)$ indicates a normal distribution with a mean of $m$ and a standard deviation of $s$. The graphical model of the proposed model is shown in Figure 4. In the figure, $J$ represents the number of learners, $R$ represents the number of reading passages, and $I_r$ represents the number of questions corresponding to the $r$-th reading passage.

The proposed model is designed such that the difficulty of a question follows a normal distribution with the difficulty of the corresponding reading passage as its mean. This allows the model to estimate the difficulty of the reading passage based on the difficulty of the questions. These parameters are estimated using the Markov chain Monte Carlo method with the No-U-Turn Sampler (Hofman & Gelman, 2014; Uto, 2021; Uto et al., 2023) based on the Hamiltonian Monte Carlo approach (Brooks et al., 2011).

### *4.2 Selecting Reading Passages with Appropriate Difficulty*

Here, we propose a method that selects reading passages appropriate to the learner's ability level, based on the difficulty of the passages estimated using the proposed IRT model, and then generates questions of suitable difficulty. As discussed in Section 3, presenting questions with difficulty $b = \theta$ is appropriate. This suggests that selecting a reading passage with $b^{(r)} = \theta$ and generating questions using the difficulty-controllable QG method with the selected reading passage, along with $b = \theta$, is a reasonable procedure.

However, in actual learning environments, the learner's ability $\theta$ is not known in advance. Therefore, this study employs the adaptive QG framework explained in Section 3. The adaptive framework, which involves selecting reading passages, generating questions, and updating learner ability, is as follows.

1. Initialize a learner's ability estimate $\hat{\theta} = 0.0$.
2. Select a reading passage with a difficulty $b^{(r)}$ close to $\hat{\theta}$.
3. Using the selected reading passage and specify the question difficulty $b = \hat{\theta}$, generate a question and answer using the difficulty-controllable QG method.
4. Present the generated question to the learner and collect their correct/incorrect responses.

5. Update the learner's ability estimate $\hat{\theta}$ from the responses given the difficulty of the presented question $b$.
6. Repeat steps 2 through 5.

By following these steps, it is possible to select reading passages and generate questions with difficulty levels that match the learner's ability.

## 5. Experiments

In this section, we describe the procedure and results of experiments conducted on a benchmark dataset to evaluate the effectiveness of the proposed method.

### *5.1 Effects of Selecting Appropriate Reading Passages*

We first conducted an experiment to evaluate whether the difficulty of the reading passages affects the difficulty of the generated questions. The experiment was conducted using the SQuAD dataset introduced in Section 2. The SQuAD dataset consists of training data and test data. In the following experiments, the training data is used for the preliminary training of the QG method and the construction of QA systems. Furthermore, a portion of the test data is used for training the difficulty-controllable QG method, along with the IRT-based difficulty estimation for the questions and reading passages, while the remaining part of the test data is used for evaluating the generation performance. The detailed experimental procedure is as follows:

1. The difficulty-controllable QG method was trained using the SQuAD training data $D^{(train)}$ without considering difficulty, as suggested by Uto et al. (2023).
2. QA systems with varying ability levels were constructed. Specifically, we used 12 BERT variants and trained them using 600, 1200, 1800, 2400, and 3000 data points randomly extracted from $D^{(train)}$. As a result, we obtained a total of 60 QA models.
3. The response data from the 60 QA systems for the questions in $D^{(eval)}$ were collected and then the difficulties of the reading passages $b^{(r)}$ and the questions $b$ within $D^{(eval)}$ were estimated using the proposed IRT model. In this experiment, we set the hyperparameter $\sigma$ of the proposed IRT model to 10.

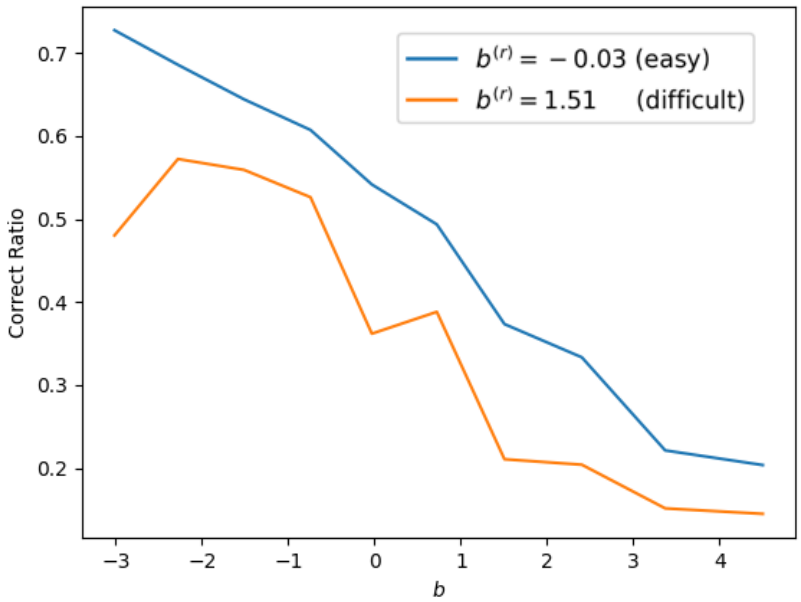


*Figure 5*. Correct Ratio of 60 QA Systems for Questions Generated from Reading Passages with Different Difficulty Values

4. After adding the estimated difficulties $b^{(r)}$ and $b$ to $D^{(eval)}$, we divided the data into 90% and 10%. The 90% of data, designated as $D_b^{(train)}$, was used to train difficulty-controllable QG method, while the remaining 10%, designated as $D_b^{(eval)}$, was used as test data.
5. Using the $D_b^{(train)}$, the difficulty-controllable QG method was trained. Then, using the trained difficulty-controllable QG method, we generated questions and answers for two reading passages in $D_b^{(eval)}$ with different difficulty levels, $b^{(r)} = -0.03$ and $1.51$ while varying the target question difficulty $b$ from $-3$ to $3$.

Figure 5 shows the results. The horizontal axis represents the question difficulty, and the vertical axis represents the correct ratio of the 60 QA systems for the generated questions. The blue line represents the results for the questions generated from a reading passage with $b^{(r)} = -0.03$, while the orange line represents those generated from a passage with $b^{(r)} = 1.51$. This figure indicates that the higher the difficulty of the reading passage, the lower the correct ratio. This suggests that the difficulty of the reading passage affects the difficulty of the generated questions, emphasizing the importance of selecting appropriate reading passages to generate questions with suitable difficulty levels.

### *5.2 Performance of Adaptive Question Generation*

Next, we conducted an experiment to evaluate the performance of adaptive QG. Specifically, we used a QA system with $\theta = 0.755$ as a target learner and investigated the adaptive QG framework with the following two strategies for selecting reading passages:

- *Proposed adaptive selection*: Selects the reading passage with the difficulty level closest to $\hat{\theta}$.
- *Random selection:* Selects a reading passage randomly.

Note that for both methods, QG was performed using the difficulty-controllable QG method, with the question difficulty set closest to $\hat{\theta}$.

Figure 6 shows the trajectories in ability estimates obtained using these methods. The horizontal axis represents the number of questions administered, with a maximum of 50, while the vertical axis represents the ability estimates $\hat{\theta}$. The blue line indicates the true ability value of $\theta = 0.755$, the orange line represents $\hat{\theta}$ obtained from the method using the proposed adaptive selection, and the green line represents $\hat{\theta}$ obtained from the method using the random selection.

From these results, it can be seen that the proposed method approaches the true ability value more efficiently. This suggests that selecting appropriate reading passages using

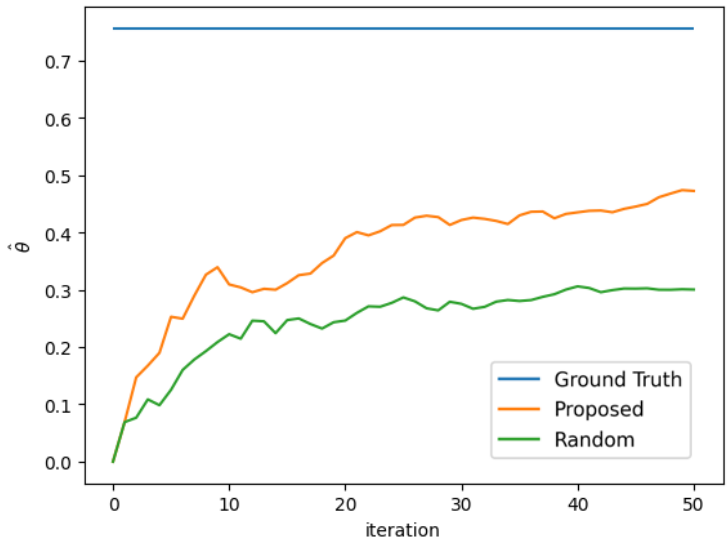


*Figure 6*. Trajectory in Ability Estimates for a QA System with Ability of 0.755

the proposed method is more efficient for estimating learner ability, thereby enhancing the selection of more appropriate difficulty values for the learner.

## 6. Conclusion

In this study, we aimed to address the limitation of existing difficulty-controllable QG methods, that is the ignorance of the difficulty of the reading passages. To this end, we first proposed a new IRT model capable of estimating the difficulty of both the questions and the reading passages by applying a hierarchical Bayesian approach to the Rasch model. We then proposed a framework based on the adaptive QG method to select reading passages and generate questions that are appropriate for each learner's reading ability. The experiments demonstrated that (1) the difficulty of the reading passages influences the difficulty controllability of QG, and (2) selecting reading passages and generating questions with difficulty levels appropriate for each learner's ability is effective for efficiently estimating learner ability and facilitating more accurate difficulty control.

There are three main challenges for future works. The first challenge involves handling new reading passages whose difficulty has not been pre-calibrated. Our adaptive QG method currently assumes to select reading passages from the training dataset because the difficulty level of each passage must be known in advance. However, reusing reading passages may affect the reliability of ability measurement. To address this issue, we will examine to integrate a new approach capable of estimating the difficulty of unseen reading passages. The second challenge pertains to model training for difficulty controllability. The difficulty-controllable QG method used in this study was trained to maximize the likelihood of question texts within the training dataset, meaning it did not directly optimize for difficulty control accuracy. To enhance this accuracy, an approach could involve additional training using reinforcement learning, which directly optimizes difficulty control accuracy. The third challenge is that it is difficult to appropriately control the difficulty for items not included in the training dataset. For example, controlling the difficulty for types of texts not included in the dataset (e.g., narrative texts) or for groups with cognitive levels different from those in the dataset is challenging. Addressing this issue will also be a task for future work.

## References


Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Cheng, Y., Li, S., Liu, B., Zhao, R., Li, S., Lin, C., & Zheng, Y. (2021). Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 5968–5978.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Du, X., Shao, J., & Cardie, C. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1342–1352.

Gao, Y., Bing, L., Chen, W., Lyu, M., & King, I. (2019). Difficulty Controllable Generation of Reading Comprehension Questions. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 4968–4974.

Goto, T., Tomikawa, Y., & Uto, M. (2024). Enhancing Diversity in Difficulty-Controllable Question Generation for Reading Comprehension via Extended T5. *The 32nd International Conference on Computers in Education.*

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. *APA educational psychology handbook, Vol. 3. Application to learning and teaching*, 451–473.

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *The Ninth International Conference on Learning Representations.*

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 609–617.

Homan, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research, 15*(1), 1593-1623.

Kurdi, G., Leo, L., Parsia, B., Sattler, U., & Al-Emari, A. (2020). A systematic review of automatic question generation for educational purposes, *International Journal of Artificial Intelligence in Education, 30*(1), 121–204.

Le, N. T., Kojiri, T., & Pinkwart, N. (2014). Automatic question generation for educational applications – The state of art. *Advanced computational methods for knowledge engineering*, 325–338.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Mostow, J., & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, 465–472.

Perkoff, E. M., Bhattacharyya, A., Cai, J., & Cao, J. (2023). Comparing neural question generation architectures for reading comprehension. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 556–566.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21*(140), 1-67.

Rathod, M., Tu, T., & Stasaski, K. (2022). Educational multi-question generation for reading comprehension. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), 216-223.*

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Subramanian, S., Wang, T., Yuan, X., Zhang, S., Trischler, A., & Bengio, Y. (2018). Neural models for key phrase extraction and question generation. *Proceedings of the Workshop on Machine Reading for Question Answering*, 78–88.

Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., & Wang, S. (2018). Answer-focused and position-aware neural question generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3930–3939.

Tomikawa, Y., & Uto, M. (2024). Difficulty-Controllable Multiple-Choice Question Generation for Reading Comprehension Using Item Response Theory. In *International Conference on Artificial Intelligence in Education*, 312-320.

Uto, M. (2021). A Multidimensional Generalized Many-Facet Rasch Model for Rubric-Based Performance Assessment. *Behaviormetrika, 48*(3), 425–457.

Uto, M., Aomi, I., Tsutsumi, E., & Ueno, M. (2023). Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies, 16*(6), 983–1000.

Uto, M., Tomikawa, Y., & Suzuki, A. (2023). Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, 119–129.

Ueno, M., & Miyazawa, Y. (2018). IRT-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies, 11*(4), 415–428.

Ueno, M., & Miyazawa, Y. (2022). Two-stage uniform adaptive testing to balance measurement accuracy and item exposure. In *International Conference on Artificial Intelligence in Education*, 626–632.

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing* (Vol. 10). Springer.

Zhou, Q., Yang, N., Wei, W., Tan, C., Bao, H. & Zhou, M. (2017). Neural Question Generation from Text: A Preliminary Study. *National China Computer Federation Conference on Natural Language Processing and Chinese Computing*, 662–671.

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