Format-Aware Item Response Theory for Predicting Vocabulary Proficiency

Boxuan Ma
OpenDNA Inc.
boxuan@open-dna.jp

Gayan Prasad Hettiarachchi
OpenDNA Inc.
gayan@open-dna.jp

Yuji Ando
OpenDNA Inc.
ando@open-dna.jp

ABSTRACT

Vocabulary proficiency testing plays a vital role in identifying the learner's level of vocabulary knowledge, which can be used to provide personalized materials and feedback in language-learning applications. Item Response Theory (IRT) is a classical method that can provide interpretable parameters, such as the learner's ability, question discrimination, and question difficulty in many language-proficiency testing environments. Many vocabulary proficiency tests include more than one type of question format. However, traditional IRT lacks the capability to tap into the information present within question texts and question formats, which can be ideally used to gauge a learner's underlying skills in more detail. In addressing this, we propose a model to reinforce traditional IRT with deep learning to exploit the information hidden within question content and format. Experimental results on a sample real-world dataset demonstrate the effectiveness of the proposed model, highlighting that question-related information can be utilized to predict a learner's performance more accurately.

Keywords

Item response theory, Deep learning, Item difficulty, English vocabulary.

1. INTRODUCTION

Vocabulary proficiency assessment plays an important part in language education and has lately gained increased popularity in online language learning. It is crucial to identify the learners' English vocabulary proficiency to higher accuracy in providing personalized materials and adaptive feedback in language-learning applications [1]. With the estimated learners' vocabulary knowledge state, systems can better gauge the attainment levels of learners and tailor the learning materials accordingly. Moreover, learners may also develop better learning plans to deal with their specific weaknesses and maximize their learning efficacy depending on the results. Most importantly, it can help place a second-language learner quickly in the ideal content space when returning to the application after a long break during which he or she may have forgotten a lot or, conversely, have progressed in the target language outside the realm of the application [2].

Computerized adaptive testing (CAT) is a mode of testing that has gained popularity because of its unparalleled ability to measure latent abilities in large-scale testing environments [3]. In CAT, estimating the difficulty level, also called item calibration, is essential for maintaining, updating, and developing new items for an item bank. Item Response Theory (IRT) [4] is a classical method widely used to determine item difficulties. IRT can predict student performance using the logistic-like item response function and provide interpretable parameters. For this reason, different IRT models have been widely applied in CAT applications [5].

Although IRT has made a great deal of success and is widely applied, some problems still limit its usefulness. The critical drawback of traditional IRT is that it can only exploit the response results and ignore the actual contents and formats of the items [6]. Thus, IRT cannot capture the rich information hidden within question texts and underlying formats. This problem leaves no possibility of generalizing item parameters to unseen items and understanding the format's impact on the difficulty of items [2]. In addition, IRT only provides an overall latent trait for learners, while each question usually assesses different knowledge concepts or skills [7]. Thus, enhancing IRT to provide detailed results on each knowledge concept or skill in a reliable way is still an open issue.

Many researchers are beginning to focus on new approaches for estimating the difficulty of questions or items to improve traditional IRT. Studies have already shown that the representational information of questions is significantly related to the difficulty level. For English vocabulary questions, word length and corpus frequency prove to be essential factors for predicting vocabulary difficulty [8], while the average word and sentence lengths have been used as key features to predict English text difficulty [9]. Along these lines, many works have begun to estimate difficulty parameters based on items' textual content using deep neural networks [2].

In vocabulary proficiency assessment, some studies have indicated that even for the same vocabulary item, different question formats impact the difficulty level and explanatory power in predicting receptive skills [10]. The ability of learners to fully comprehend a specific word can be divided into different components. The best-known and most widely used framework is Nation's division of vocabulary knowledge into nine components of 'word knowledge' (e.g., spelling, word parts, meaning, grammatical functions, and collocation) [11]. The framework has been instrumental in describing the totality of what learners need to know. However, no single question format can adequately describe vocabulary comprehension. Usually, different question formats are used to assess different skills, such as learners' reading, writing, listening, and speaking skills collectively. However, IRT only provides an overall latent trait on the question level and cannot provide more detailed results on the underlying skills.

B. Ma, G. P. Hettiarachchi, and Y. Ando. Format-aware item response theory for predicting vocabulary proficiency. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 695–700, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853091

2. RELATED WORK

IRT is one of the most time-tested theories for estimating latent abilities and has been used in educational testing environments since the 1950s [12]. There are several IRT models widely in use, such as the 1-parameter, 2-parameter, and 3-parameter models [13,14,15]. Extended from IRT, Multidimensional Item Response Theory (MIRT) [17] tries to meet multidimensional data demands by including an individual's multidimensional latent abilities for each skill. Although MIRT goes a step further to include the knowledge-concept proficiencies of individuals, it is sensitive to the knowledge concepts on which they have high latent abilities [17]. In addition, since the process of estimating the parameters for MIRT is the same as IRT, these two models share the same shortcomings.

With the recent surge in interest in deep learning, many works have begun to incorporate deep learning models into IRT to address the shortcomings of traditional IRT models. For example, the synthesis of Bayesian knowledge tracing (BKT) and IRT [18, 19] empowers the individualization of questions and learners. Recently, Deep-IRT [20] was proposed by combining a dynamic key-value memory network (DKVMN) [21] with an IRT module to improve the explanatory capabilities of the parameters. Furthermore, Emiko et al. [22] improved Deep-IRT with two independent neural networks for students and items.

Other IRT-based works have focused on improving the estimation accuracy of parameters by exploiting the semantic representations from question texts. Cheng and Liu [23] proposed a general Deep Item Response Theory (DIRT) framework that uses deep learning to estimate item discrimination and difficulty parameters by extracting information from item texts. Benedetto et al. [24] adopted transfer learning on Transformer language models [25] and performed the estimation of the difficulty parameter. Hsu et al. [26] proposed a method for automated estimation of multiple-choice items' difficulty for social studies tests. Their findings suggest that the semantic similarity between a stem and the options strongly impacts item difficulty. Susanti et al. [27] proposed a system for automatically generating questions for vocabulary tests. Factors such as the reading passage difficulty, semantic similarity between the correct answer and distractors, and distractor word difficulty level are all considered for controlling generated items' difficulty in this system.

Studies looking into language tests also tried to predict the item difficulty and automatically generate items of various difficulty levels. Many of these studies have investigated the relationship between test item difficulty and linguistic features such as passage length, word length, and word frequency. Hoshino and Nakagawa [28] used a support vector machine to estimate the difficulty of cloze items for a CAT. Beinborn et al. [29] used Natural Language Processing (NLP) to predict c-test difficulty at the word-gap level, using a combination of factors such as phonetic difficulty and text complexity. Loukina et al. [30] conducted a study to investigate which textual properties of a question affect the difficulty of listening items in an English language test. Settles et al. [31] used Machine Learning and NLP to induce proficiency scales and then used linguistic models to estimate item difficulty directly for CAT. However, these studies did not consider a variety of item formats that would typically appear in a test, and failed to consider linguistic skills in vocabulary learning. Recently, Brian and Andrew [48] have incorporated rich linguistic features (lexical, morphological, and syntactic features) as skills in skill-based models for learners' vocabulary learning performance prediction. It highlighted that the use of linguistic skills is quite helpful in this regard. However, their work also failed to consider question formats' influence on the difficulty level and different receptive skills. In addressing this, here we incorporate item-format information and associated skill requirements to improve the estimations of IRT parameters, and in effect, the prediction accuracy of a learners' performance.

3. PROPOSED METHOD

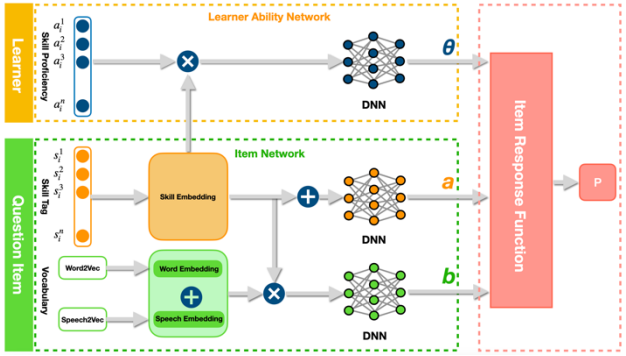

Figure 1. Overview of the proposed framework.

3.1 Framework

Inspired by previous studies [6, 23], we propose a framework to enhance traditional IRT with deep learning, which aims to obtain the learner parameter (ability) and item parameters (discrimination and difficulty) to predict learner performance in vocabulary questions. In achieving this, as shown in Figure 1, our framework comes with three parts: the leaner ability network, item network, and prediction module.

3.1.1 Item Network

The items' characteristics, i.e., the difficulty and discrimination parameters, are calculated in the item network.

For a vocabulary question presented in a specific format, two elements influence the item's characteristics: the target vocabulary and the required skills to respond correctly. For the target vocabulary, the semantic features are embedded into a d-dimensional vector v using pre-trained Word2Vec [32] vector v_w and Speech2Vec [33] vector v_s, where $v = v_w \oplus v_s$ and d = 50. And the required skills for this question are represented by one-hot vectors $S = (S_1, S_2, \dots, S_n), S_i \in \{0,1\}^n$, where n is the number of required skills. Then, we utilize a d-dimensional dense layer to acquire the dense embedding for each skill S_i for training, the dense embedding of S_i as s_i, and $s_i \in \mathbb{R}^d$:

$$s_i = S_i W_s , \quad (1)$$

where $W_s \in \mathbb{R}^{p \times d}$ are the parameters of the dense layer. Then we use the target vocabulary embedding and the skill embedding to obtain the item parameters (discrimination and difficulty).

Discrimination. Question discrimination a can be used to analyze learner performance distribution on the question. Inspired by previous works [17, 23], we learn a from required skills that correspond to the question. A deep neural network (DNN) is trained to estimate a. Specifically, we sum the dense embedding of required skills to get a d-dimensional vector $A \in \mathbb{R}^d$. Then, we input A into the DNN to estimate a. Finally, we normalize a so that it is in the range $[-4, 4]$ [16]. The definition of a is as follows:

$$a = 8 \times (sigmoid(DNN_a(A)) - 0.5), \; A = s \oplus s . \quad (2)$$

Difficulty. Question difficulty b determines how hard the question is. Adopting from previous works [8, 9], we predict b based on the semantic features of the target word. In addition, the depth and width of the required skills examined by the question also significantly impact the difficulty. The deeper and broader the required skills being examined, the more difficult the question is [23]. Therefore, we adopt a DNN to model b based on the target vocabulary embedding v and the required skills s depending on the corresponding question format. Like the discrimination, we normalize b so that it is in the range [−4, 4] [16]. The definition of b is as follows:

$$b = 8 \times (sigmoid(DNN_b(B)) - 0.5), \quad B = s \odot v\,. \quad (3)$$

3.1.2 Learner Ability Network

In the learner network, the proposed method calculates a learner's ability. For a learner, we initialize a proficiency vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ randomly, where $\alpha_i \in [0, 1]$ represents the degree of proficiency of a learner on a specific skill i.

Ability. Learners' ability θ has strong interpretability for their performance on questions. It is closely related to the proficiency of various skills tested in the questions [23]. Therefore, we multiply the corresponding proficiency α with the skill dense embedding of the questions s and get a d-dimensional vector $\Theta \in \mathbb{R}^d$. Then we input Θ into a DNN to learn the ability parameter, which is defined as follows:

$$\theta = DNN\theta\,(\Theta), \; \Theta = \alpha \odot s\,. \quad (4)$$

3.1.3 Prediction of Learner Response

Like previous works [20, 22, 23], the proposed method predicts a learner's response performance to a question as a probability. We input the trained parameters, namely, θ, a, and b into the item response function (Eq.5) to predict $P(\theta)$, the learner's probability of answering the specific question correctly.

$$P(\theta) = \frac{1}{1+e^{-a(\theta - b)}}. \quad (5)$$

3.1.4 Model Learning

The parameters to be updated in the proposed framework mainly exist in two parts: learner ability network and item network. The updating parameters include the proficiency vector α and skill dense embedding weights W_s. In addition, the weights of the three DNNs $\{W_{DNN_a}, W_{DNN_b}, W_{DNN_\theta}\}$ are updated as well.

The loss function of the proposed method is the negative log-likelihood function. The learner's response is recorded as 1 when he/she answers the item correctly and 0 otherwise. For learner i and question j, let r_{ij} be the actual score for learner i on question j, and $\tilde{r}_{ij}$ be the predicted score. Thus, the loss for learner i on question j is defined as:

$$\mathcal{L} = r_{ij} log\tilde{r}_{ij} + (1 - r_{ij})log(1 - \tilde{r}_{ij}). \quad (6)$$

Using Adam optimization [34], all parameters are learned simultaneously by directly minimizing the objective function.

4. EVALUATION

4.1 Dataset

Our real-world dataset came from one of Japan's most popular English-language learning applications. We tentatively used a sample dataset from 129 application users who newly registered in 2021, and most of them are Japanese students learning English. This dataset included 1,900 English words labeled by the Common European Framework of Reference for Languages (CEFR), mainly in the B1/B2 range. Each word in the dataset had six different question types collectively assessing reading, writing, listening, and speaking skills. The dataset included the initial responses (when encountering for the first time) of the users to such questions.

4.1.1 Item Formats

The knowledge pertaining to English words is not all-or-none as with the case with any other language. Rather, there are different aspects, such as knowledge of the reading, writing, listening, speaking, grammatical behavior, collocation behavior, word frequency, stylistic register constraints, conceptual meaning, the associations a word has with other related words, and so on [11, 35]. Hence, as summarized in Table 1, there are six different question formats to collectively assess reading, writing, listening, and speaking skills of vocabulary learning in our dataset. For each format, we indicate the linguistic skill(s) required to tackle the question (L = listening, R = reading, S = speaking, W = writing) and some of the evidence from the literature supporting this assignment. Below are the descriptions of the six question formats. Multiple-choice definition: choose the Japanese description of the English word. Multiple-choice recall: choose the corresponding English word given the Japanese description. Spelling: type the spelling of the English word given the Japanese description. Cloze test with spelling: type in the blank with the appropriate English word. Multiple-choice listening: choose the corresponding English word given the pronunciation. Multiple-choice cloze test: choose the appropriate English word to fill the blank.

Table 1. Summary of question formats and required skill(s).

Label	Question Format	Skills	References
Format 1	Multiple-choice definition	R	[40,41,44]
Format 2	Multiple-choice recall	R	[40,41]
Format 3	Spelling	R S W	[39]
Format 4	Cloze test with spelling	R S W	[39,42]
Format 5	Multiple-choice listening	L R	[38,43]
Format 6	Multiple-choice cloze test	R W	[31,39,42]

4.2 Experimental Settings

We conducted extensive experiments to evaluate the accuracy of our model in predicting the performance of learners and compared it with several existing models. To set up the experiments, we partitioned the dataset, where the question-user interactions were divided into the training and testing sets at different ratios: 60%, 70%, 80%, and 90%.

We name our method Format-Aware IRT(FIRT). FIRT-S is a variant of FIRT, which only uses speech embedding based on the Speech2Vec, and FIRT-W is a variant that only uses word embedding based on the Word2Vec. We compared our method's performance with IRT, MIRT, and Probabilistic matrix factorization (PMF) [36].

Following previous works [23, 37], we chose four widely used metrics for the evaluation: Prediction Accuracy (ACC), Area Under Curve (AUC), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The smaller the values of RMSE and MAE, and the larger the values of AUC and ACC, the better the results.

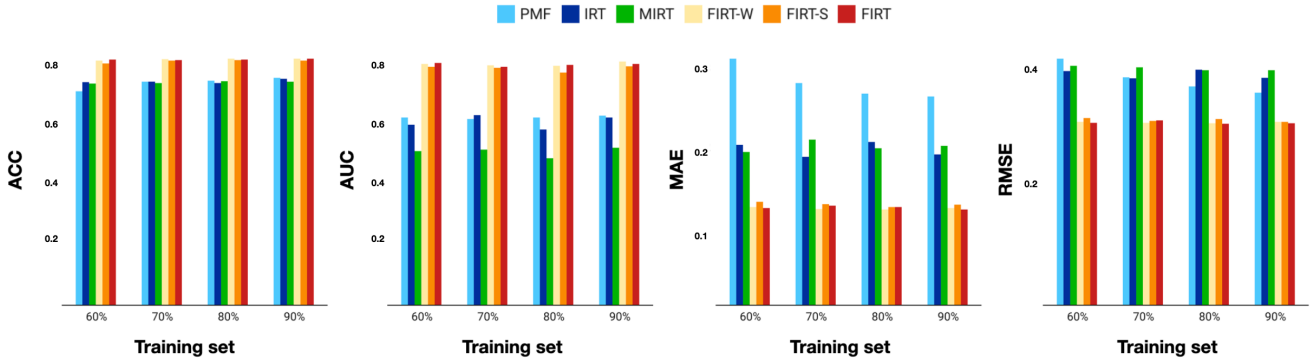

Figure 2. Comparison of learner performance prediction among different methods.

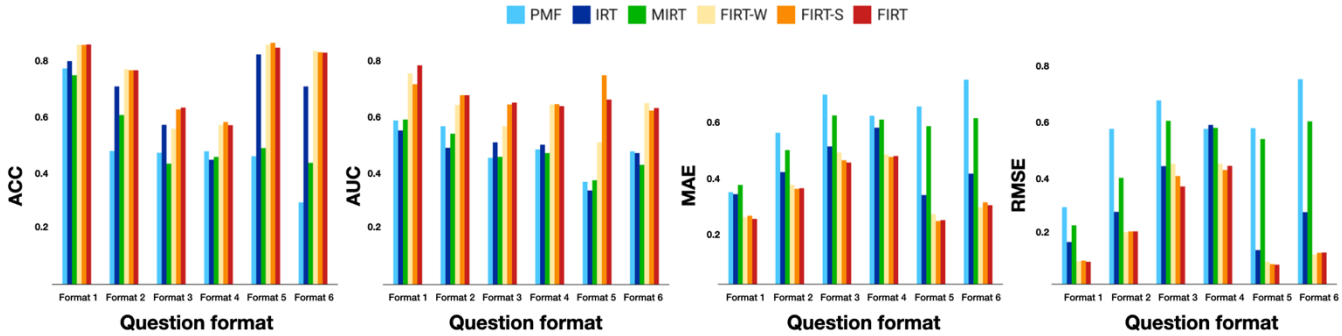

Figure 3. Comparison among different question formats.

4.3 Results

4.3.1 Performance Prediction

The overall results on all four metrics are shown in Figure 2 for six different models predicting learners' performance. We observe that our proposed models (FIRT, FIRT-S, and FIRT-W) perform better than other traditional baseline models, such as IRT, MIRT, and PMF. It is clear that our deep learning-based models can effectively make use of the vocabulary content and format information to improve the performance.

4.3.2 Impact of Different Formats

Many studies predicting vocabulary knowledge use only a single question format. However, results based on a single format can be misleading because it might be gauging a limited skill space pertaining to vocabulary knowledge. Understanding the differences among various formats may provide insight for developing tests and tools that can measure proficiency on a much-balanced scale [47]. Along this line, we conducted experiments to evaluate our model's ability to predict performance on different question formats, and to illustrate the variability of performance depending on the format.

The results in Figure 3 show that our models perform better in all question formats. This indicates that the vocabulary content and format information, together with the underlying skill proficiency, help predict learners' performance better, which are typically ignored in the traditional methods. Also, we observe that the prediction performance is strongly affected by the question format. As we mentioned earlier, different question formats assess different linguistic skills like reading, writing, listening, and speaking. The results show that all models have considerably satisfactory performance for multiple-choice items, which assess only one or two underlying skills and are easy to understand and answer [45, 46]. However, the prediction performance for Format 3 (Spelling) and Format 4 (Cloze test with spelling) are deficient compared with others, which intuitively suggests that responses to question formats that necessitate multiple underlying skills are more difficult to predict accurately. In addition, we noticed that FIRT-S (using speech embedding) performs slightly better than other models for Format 5 (Multiple-choice listening). This implies that using other features besides semantic features may improve the performance. Moreover, the findings implies that testers should consider the effect of different formats when assessing vocabulary knowledge and strive to use a combination of formats in vocabulary assessment to gauge a broader skill space.

5. CONCLUSION

In this work, we proposed a framework that reinforces IRT with deep learning routines that take full advantage of the questions' representational information, such as the question contents, formats, and the required linguistic skill(s) to tackle the question. Experiments were conducted to confirm the effectiveness of the proposed approach, and the results showed that our method performs better than other methods. We highlight that vocabulary content and format information together with the required skill set is useful in accurately predicting learners' vocabulary proficiency.

However, there are some limitations in this work. The dataset is relatively small, and the learner base is limited to learners of the same language background. For future work, we plan to collect more data on learners of various backgrounds, which may be useful when generalizing the method to a broader audience. Also, it is likely that the six item formats explored in this work over-index on language reception skills rather than production skills (i.e., writing and speaking). In going forward, we need to test more writing and speaking questions, and include additional linguistic skills to expand the capabilities of our model.

6. REFERENCES

[1] Avdiu, D. and Bui, V. 2019. Predicting learner knowledge of ind dividual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning,* pages 1-9.

[2] Robertson, F. 2021. Word Discriminations for Vocabulary Inventory Prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1188-1195.

[3] Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. 2000. *Computerized adaptive testing: A primer*. Routledge.

[4] Embretson, Susan E., and Steven P. 2013. Reise. *Item response theory*. Psychology Press.

[5] Hambleton, R. K. 1989. Principles and selected applications of item response theory.

[6] Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., and Hu, G. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 100-115.

[7] Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., ... and Hu, G. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 1352-1359.

[8] Culligan, B. 2015. A comparison of three test formats to assess word difficulty. *Language Testing*, *32*(4), 503-520.

[9] Beinborn, L., Zesch, T., and Gurevych, I. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517-530.

[10] Kremmel, B., and Schmitt, N. 2016. Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words?. *Language Assessment Quarterly*, *13*(4), 377-392.

[11] Nation, I. S. 2001. Learning vocabulary in another language. Cambridge university press.

[12] Embretson, S. E., and Reise, S. P. 2013. Item response theory. Psychology Press.

[13] Baker, F. B., & Kim, S. H. 2004. Item Response Theory: Parameter Estimation Techniques, Second Edition. *Statistics: A Series of Textbooks and Monographs*. Taylor & Francis.

[14] Lord, F. M., and Novick, M. R.1968. Statistical Theories of Mental Test Scores. Addison-Wesley.

[15] Ueno, M., and Miyasawa, Y. 2015. Probability based scaffolding system with fading. *Artificial Intelligence in Education. 17th International Conference*, AIED, pages 237–246.

[16] Van der Linden, W. J., and Hambleton, R. K. (1997). Handbook of item response theory. *Taylor & Francis Group. Citado na pág*, *1*(7), 8.

[17] Yao, L., & Schwarz, R. D. 2006. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied psychological measurement*, *30*(6), 469-492.

[18] Mohammad M. Khajah, Yun Huang, Jos´e P. Gonz´alez-Brenes, Michael C. Mozer, and Peter Brusilovsky. 2014. Integrating knowledge tracing and item response theory: A tale of two frameworks. In *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environment*, pages 7–15.

[19] Kevin H. Wilson, Yan Karklin, Bojian Han, and Chaitanya Ekanadham. 2016. Back to the basics: Bayesian ex- tensions of IRT outperform neural networks for proficiency estimation. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 539–544.

[20] C. Yeung. 2019. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. In *Proceedings of the 12th International Conference on Educational Data Mining*, pages 683-686.

[21] Zhang, J., Shi, X., King, I., and Yeung, D. Y. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*. pages 765-774.

[22] Tsutsumi, E., Kinoshita, R., and Ueno, M. 2021. Deep-IRT with Independent Student and Item Networks. In *Proceedings of the 14th International Conference on Educational Data Mining*, pages 510-517.

[23] Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., and Hu, G. 2019. DIRT: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. pages 2397-2400.

[24] Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., and Turrin, R. 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.

[25] Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 2978-2988.

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

[27] Susanti, Y., Nishikawa, H., Tokunaga, T., and Hiroyuki, O. 2016. Item Difficulty Analysis of English Vocabulary Questions. In *International Conference on Computer Supported Education*. Vol. 2, pages. 267-274.

[28] Hoshino, A., and Nakagawa, H. 2010. Predicting the difficulty of multiple-choice close questions for computer-adaptive testing. *Natural Language Processing and its Applications*, 46, pages 279-292.

[29] Beinborn, L., Zesch, T., and Gurevych, I. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517-530.

[30] Loukina, A., Yoon, S. Y., Sakano, J., Wei, Y., and Sheehan, K. 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.

[31] Settles, B., T LaFlair, G., and Hagiwara, M. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, *8*, 247-263.

[32] Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

[33] Chung, Y. A., and Glass, J. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.

[34] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[35] Dale, E. 1965. Vocabulary measurement: Techniques and major findings. *Elementary english*, *42*(8), 895-948.

[36] Mnih, A., and Salakhutdinov, R. R. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems*, *20*.

[37] Wu, R., Xu, G., Chen, E., Liu, Q., and Ng, W. 2017. Knowledge or gaming? Cognitive modelling based on multiple-attempt response. In *Proceedings of the 26th International Conference on World Wide Web Companion*. pages 321-329.

[38] Bradlow, A. R., and Bent, T. 2002. The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1), 272-284.

[39] Khodadady, E. 2014. Construct validity of C-Tests: A factorial approach. *Journal of Language Teaching and Research*, 5(6), 1353.

[40] Staehr, L. S. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139-152.

[41] Milton, J. 2010. The development of vocabulary breadth across the CEFR levels. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 211-232.

[42] Klein-Braley, C. 1997. C-Tests in the context of reduced redundancy testing: An appraisal. *Language testing*, *14*(1), pages 47-84.

[43] Milton, J., Wade, J., and Hopkins, N. 2010. Aural word recognition and oral competence in English as a foreign language. *Insights into non-native vocabulary teaching and learning*, 83-98.

[44] Brown, J., Frishkoff, G., and Eskenazi, M. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.* pages 819-826.

[45] Kremmel, B., and Schmitt, N. 2016. Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words?. *Language Assessment Quarterly*, *13*(4), 377-392.

[46] Kilickaya, F. (2019). Assessing L2 vocabulary through multiple-choice, matching, gap-fill, and word formation items. *Lublin Studies in Modern Languages and Literature*, 43(3), 155-166.

[47] Bowles, R. P., and Salthouse, T. A. 2008. Vocabulary test format and differential relations to age. *Psychology and Aging*, 23(2), 366.

[48] Zylich, B., and Lan, A. 2021. Linguistic Skill Modeling for Second Language Acquisition. In *LAK21: 11th International Learning Analytics and Knowledge Conference.* pages 141-150.