Exploring the Effectiveness of Vocabulary Proficiency Diagnosis Using Linguistic Concept and Skill Modeling

Boxuan Ma
Kyushu University
OpenDNA Inc.
boxuan@artsci.kyushu-u.ac.jp

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

Sora Fukui
OpenDNA Inc.
Tokyo, Japan
fukui@open-dna.jp

Yuji Ando
OpenDNA Inc.
Tokyo, Japan
ando@open-dna.jp

ABSTRACT

Vocabulary proficiency diagnosis plays an important role in the field of language learning, which aims to identify the level of vocabulary knowledge of a learner through his or her learning process periodically, and can be used to provide personalized materials and feedback in language-learning applications. Traditional approaches are widely applied for modeling knowledge in science or mathematics, where skills or knowledge concepts are well-defined and easy to associate with each item. However, only a handful of works focus on defining knowledge concepts and skills using linguistic characteristics for language knowledge proficiency diagnosis. In addressing this, we propose a framework for vocabulary proficiency diagnosis based on neural networks. Specifically, we propose a series of methods based on our framework that uses different linguistic features to define skills and knowledge concepts in the context of the language learning task. Experimental results on a real-world second-language learning dataset demonstrate the effectiveness and interpretability of our framework. We also provide empirical evidence with ablation testing to prove that our knowledge concept and skill definitions are reasonable and critical to the performance of our model.

Keywords

Deep learning, Cognitive diagnosis, Vocabulary proficiency, Linguistic skill

1. INTRODUCTION

Vocabulary proficiency diagnosis is one of the key fundamental technologies supporting language education and has lately gained increased popularity in online language learning. It is crucial to identify the learners' latent proficiency level on different knowledge concepts (e.g., words) to higher accuracy in providing personalized materials and adaptive feedback in language-learning applications [1]. In practice, with the diagnostic results, systems can provide further support, such as learning planning, learning material recommendation, and computerized adaptive testing accordingly. Most importantly, it can help second-language learners to place themselves in the correct learning space or level after a long gap without using the application, during which they might have forgotten a lot or, conversely, have advanced in the target language without the use of the application [25].

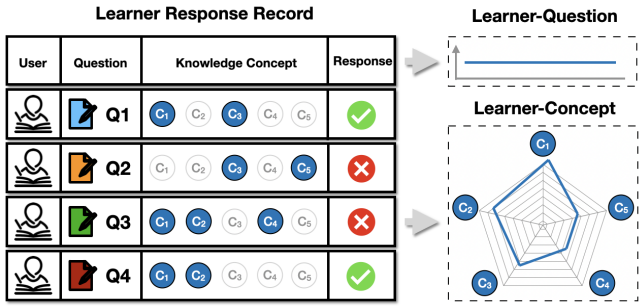

Figure 1: An example of cognitive diagnosis.

Many cognitive diagnosis methods have been proposed for knowledge proficiency diagnosis of learners. Figure 1 shows a simple example of a cognitive diagnosis system, which consists of learners, question items, knowledge concepts, and learner responses (scores). Specifically, a learner interacts with a set of questions and leaves their responses. Moreover, human experts usually label each question item with several knowledge concepts. Then, the goal is to infer their actual knowledge proficiency based on the interactions. Therefore, a cognitive diagnosis system can be abstracted as a learner-question-concept interaction modeling problem, and most previous works focus on learner-question interaction models or learner-concept interaction models [11]. For example, traditional methods like Item Response Theory (IRT) [9], Multidimensional IRT (MIRT) [24], and Matrix Factorization (MF) [23] try to model the learner-question interaction

B. Ma, G. P. Hettiarachchi, S. Fukui, and Y. Ando. Exploring the effectiveness of vocabulary proficiency diagnosis using linguistic concept and skill modeling. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 149–159, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115675

and provide learner latent traits (e.g., ability level) and the question features (e.g., difficulty level). In addition, MIRT and MF cannot provide explainable traits and IRT only provides an overall latent trait for learners, while each question usually assesses different knowledge concepts or skills. Other works such as Deterministic Inputs, Noisy-And gate (DINA) [6] try to build the learner-concept interaction instead of learner-question interaction. Unlike learner-question interaction models, learner-concept interaction models could infer the learner's traits in detail for each knowledge concept contained in the question item, despite leaving information of questions underexploited by simply replacing them with their corresponding concepts. Although great successes have been made, there are some limitations of traditional methods, which decay their effectiveness. Also, these approaches are widely applied for modeling knowledge in science or mathematics and ignore characteristics of language learning, which make it a significant research challenge to infer the mastery level of learners' vocabulary proficiency.

A critical drawback of traditional methods is that they can only exploit the response results and ignore the actual contents and formats of the items and cannot effectively utilize the rich information hidden within question texts and underlying formats [18]. Most traditional methods were proposed for scale-based tests, where a group of examinees is tested using the same small set of questions, and each examinee is supposed to respond to every question. As a result, the response data is complete and usually not large. While for learning applications nowadays, the data might be collected via different scenes, such as offline examinations and online self-regulated learning, and the distribution of response data can be of high volume but very sparse due to the large total number of items and limited questions attempted by the learners [33]. Therefore, neglecting contents and formats leaves traditional methods no possibility to utilize the relationships of different items, hence they are unable to generalize item parameters to unseen items [25]. Previous studies have already shown that the information of questions is significantly related to item parameters, for instance, the difficulty level. For language vocabulary questions, character length and corpus frequency prove to be essential factors for predicting vocabulary difficulty [5], while the average word and sentence lengths have been used as key features to predict text difficulty [2, 25]. Also, studies have indicated that different question formats impact the difficulty level and explanatory power in predicting receptive skills [16]. For the same vocabulary, different question formats are often used collectively to assess different skills, such as reading, writing, listening, and speaking skills, and many assessments have a mixture of item types. Consequently, it is important to consider the format information of the items and their influence on different traits when building a vocabulary proficiency diagnosis model.

Another important challenge is to define and use linguistic skills for vocabulary proficiency diagnosis. Although many approaches are widely applied for proficiency diagnosis, they have not frequently been applied to data generated in language learning settings. Instead, they have been primarily applied to science, engineering, and mathematics learning contexts, where skills or knowledge concepts are well-defined and easy to associate with each item. Most works use manually labeled Q-matrix to represent the knowledge relevancies of each question. For example, a math question: $6 \times 9 + 3 = ()$ examines the mastery of two knowledge concepts: Addition and Multiplication. Thus, the Q-matrix for this question could be labeled as $(1, 1, 0, ..., 0)$, where the first two positions show this question test Addition and Multiplication concepts, and other positions are labeled with zero, indicating other knowledge concepts are not included. However, proficiency diagnosis in the realm of language learning is different from other domains since linguistic skills are hard to define and need to be well-designed [21, 38].

To address these challenges, which have not been well explored in the research community, in this paper, we propose a framework for vocabulary proficiency diagnosis, which could capture the learner-question interactions more accurately using neural networks. In addition, we use linguistic features of words such as morphological and semantic features to define knowledge concepts and skills related to vocabulary and grammar knowledge that is shared between words. Extensive experimental results on a real-world second-language learning dataset demonstrate the effectiveness and interpretational power of our proposed framework. We also provide empirical evidence with ablation testing to prove that our knowledge concept and skill definitions are reasonable and critical to the performance of our model. The results show that using linguistic features to refine knowledge concepts and skills improves performance over the basic word-level model. We also explore the relationship of the question format, and in turn, its effect on the vocabulary proficiency diagnosis.

2. RELATED WORK

2.1 Cognitive Diagnosis

Cognitive diagnosis is a fundamental and important task, and many classical cognitive diagnosis models have been developed in educational psychology, such as IRT, MIRT, and DINA. IRT [9] is a widely used method and has been applied in educational testing environments since the 1950s [9]. It applies the logistic-like item response function and provides interpretable parameters. In its simplest form, IRT could be written as:

$$P(X_{ij} = 1) = \sigma(\theta_i - \beta_j),$$

where P is the probability of the learner i answering the item j correctly, σ is a logistic-like function, θ and β are unidimensional and continuous latent traits, indicating learner ability and item difficulty, respectively. Besides the basic IRT, other IRT models extend the basic one by factoring in other parameters, such as the item discrimination or guessing parameter.

IRT has proven to be a robust model. However, a single ability dimension is sometimes insufficient to capture the relevant variation in human responses. By extending the trait features into multidimensions, Reckase et al. [24] proposed MIRT, which tries to meet multidimensional data demands by including an individual's multidimensional latent abilities for each skill. MIRT goes a step further compared to IRT, however, as the process of estimating the parameters for MIRT is the same as IRT, these two models share the same shortcomings [4]. Also, latent trait vectors provided by

IRT and MIRT is not explainable enough to guide learners' self-assessment [34].

By characterizing learner features (e.g., ability) and item features (e.g., difficulty), IRT builds learner-question interaction and provides an overall latent trait for learners. However, real-world questions usually assess different knowledge concepts or skills, and an overall trait result is insufficient [20]. To provide detailed results on each knowledge concept or skill, other works try to directly build learner-concept interaction. For example, DINA [6] model the learner-concept interaction by mapping questions to corresponding concepts/ skills directly with Q-matrix, which indicates whether the knowledge concept is required to solve the question. Different from IRT, θ and β are multi-dimensional and binary in DINA, where β came directly from Q-matrix. Another two parameters, guessing g and slipping s, are also taken into consideration. The DINA formula is written as:

$$P(X_{ij} = 1) = {g_j}^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}, \quad \eta_{ij} = \prod_{k=1}^{K} {\theta_{ik}}^{\beta_{jk}},$$

where the latent response variable η_{ij} indicates whether the learner has mastered all the required knowledge to solve the question. And the probability of the learner i correctly answering item j is modeled as the compound probability that the learner has mastered all the skills required by the question without slip, and the learner does not master all the required skills but makes a successful guess. Although DINA has made great progress and shows its advantage compared to IRT in specific scenarios, it ignores the features of questions and simply replaces them with the corresponding knowledge concepts/skills, thus leaving useful information from questions underexploited.

2.2 Matrix Factorization

Besides the traditional models, the other line of studies has demonstrated the effectiveness of MF for predicting learner performance by factorizing the score matrix, which was originally widely used in the field of recommendation systems [3]. Studies have shown that predicting learner performance can be treated as a rating prediction problem since *learner*, *question*, and *response* can correspond to *user*, *item*, and *rating* in recommendation systems, respectively.

Toscher et al. [30] applied several recommendation techniques in the educational context, such as Collaborative Filtering (CF) and MF, and compared them with traditional regression methods for predicting learner performance. Along this line, ThaiNghe et al. [28] proposed multi-relational factorization models to exploit multiple data relationships to improve the prediction results in intelligent tutoring systems. In addition, Desmarais [8] used Non-negative Matrix Factorization (NMF) to map question items to skills, and the resulting factorization allows a straightforward interpretation in terms of a Q-matrix. Similarly, Sun et al. [27] proposed a method that uses Boolean Matrix Factorization (BMF) to map items into latent skills based on learners' responses. Wang et al. [36] proposed a Variational Inference Factor Analysis framework (VarFA) and utilized variational inference to estimate learners' mastery level of each knowledge concept.

Despite their effectiveness in predicting learner performance, the latent trait vectors in MF are not interpretable for cognitive diagnosis, i.e., there is no clear correspondence between elements in trait vectors and specific knowledge concepts. Also, these works have considered only learners and question items, and ignored other information that may also be useful.

2.3 Deep-learning based models

With the recent surge in interest in deep learning, many works have begun to use deep learning to address some of the shortcomings of traditional cognitive diagnosis models [13, 19, 29].

Traditional methods are often based on simple linear functions, such as the logistic-like function in IRT or the inner product in matrix factorization, which may not be sufficient. To improve precision and interpretability, some previous works focus on interaction function design and use neural networks to learn more complex non-linear functions. For example, Wang et al. [33] propose a Neural Cognitive Diagnosis (NCD) framework for Intelligent Education Systems, which leverages neural networks to automatically learn the interaction function.

Some researchers focus on incorporating the content representation from question texts into the model by neural networks, which is difficult with traditional methods. Cheng and Liu [4] 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. Wang et al. [34] applied neural networks to extract two typical types of information in the question text: knowledge concepts and extra text-related factors. Their results indicated that using such content information benefited the model and significantly improved its performance.

Other deep-learning models try to incorporate dependency relations among knowledge concepts for enhancing diagnosis performance. For example, Wang et al. [35] proposed a model based on neural networks and aggregate knowledge relationships by converting all knowledge concepts into a graph structure. Ma et al. [22] proposed the Prerequisite Attention model for Knowledge Proficiency (PAKP) to explore the prerequisite relation among knowledge concepts and use it for inferring knowledge proficiency. Recent work proposed the Relation map driven Cognitive Diagnosis (RCD) [11] model by comprehensively modeling the learner-question interactions and question-concept relations. Their model achieved better performance compared to traditional works that consider only learner-question interactions (e.g., IRT) or only question-concept interactions (e.g., DINA).

Although deep learning models have been widely explored nowadays, they have been primarily applied to learning contexts such as math, algebra, or science, where skills or knowledge concepts are well-defined and easily associated with each item. Therefore, these methods cannot be directly used in the language learning area, and linguistic skills need to be well-defined and well-designed for language proficiency diagnosis. In addition, except for the work by wang et al. [34], other aforementioned works failed to consider question for-

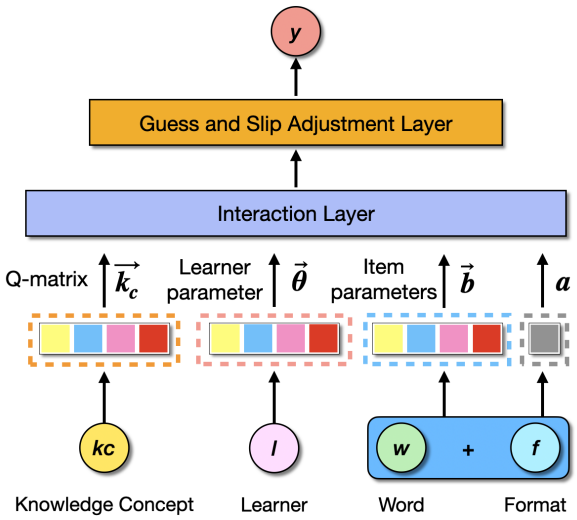

Figure 2: Overview of the proposed framework.

mats, which are important for language-learning questions and may have a significant influence on the question difficulty level and learner's performance.

3. PROPOSED METHOD

We first give the definition of our problem in Section 3.1. Then we present our proposed framework in Section 3.2.

3.1 Problem Formulation

Like every test, there are two basic elements: *user* and *item*, where a user represents a learner, and an item represents a question. We use L to denote a set of learners, Q to denote a set of questions and s to denote the learner-question interaction score. Learner question records are represented by $R = \{(l, q, s)|\, l \in L, q \in Q, s \in \{0, 1\}\}$, which means learner l responded to question q and received the score s. Each score s is in $\{0, 1\}$ where 1 indicates the question is correctly answered while 0 stands in the opposite.

Given enough question-records data R of learners, our goal is to build a model to mine learners' proficiency through the task of performance prediction.

3.2 Framework

Generally, for a cognitive diagnostic system, there are three parts that need to be considered: learner, question item, and interaction function. As shown in Figure 2, we propose a cognitive diagnostic framework with deep learning, which aims to obtain the learner parameter (proficiency) and item parameters (discrimination and difficulty). Specifically, for each response log, we use one-hot vectors of the corresponding learner and question as input and obtain the diagnostic parameters of the learner and question. Then the model learns the interaction function among the learner and item parameters and outputs the probability of correctly answering the question. After training, we get the learner's proficiency vectors as diagnostic results.

3.2.1 Item Parameters

The item's characteristics are calculated in the item network to represent the traits of a specific item. Two parameters extended from the Two-Parameter Logistic IRT model [32] are used in our model, i.e., discrimination and difficulty. The discrimination $a \in (0, 1)$ indicates the ability of an item to differentiate among learners whose knowledge mastery is high from those with low knowledge mastery, and difficulty $\boldsymbol{b} \in (0, 1)^{1\times K}$ indicates the difficulty of each knowledge concept examined by the question, where K is the number of knowledge concepts.

As we mentioned before, two elements influence the item's characteristics for a vocabulary question: the target word and the specific item format. Then the item is represented by integrating the one-hot word embedding vector $\boldsymbol{w}$ and one-hot item format embedding $\boldsymbol{f}$.

$$\boldsymbol{i} = \boldsymbol{w} \oplus \boldsymbol{f}, \tag{1}$$

where $\oplus$ is the concatenation operation. After obtaining item representation using the word embedding and item format, we input it into two different networks to estimate the question discrimination a and knowledge difficulty $\boldsymbol{b}$. Specifically:

$$a = \sigma(F_a(\boldsymbol{i})), \tag{2}$$

$$\boldsymbol{b} = \sigma(F_b(\boldsymbol{i})) \tag{3}$$

Where F_a and F_b are discrimination and difficulty networks, respectively, and σ is the sigmoid function.

3.2.2 Learner Parameter

In the learner network, the proposed method characterizes the traits of learners, which is closely related to the proficiency of various knowledge concepts or skills tested in the question and would affect the learner's performance. Specifically, each learner is represented with a proficiency vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$, where $\theta_i \in [0, 1]$ represents the degree of proficiency of a learner on a specific knowledge concept or skill i and the goal of our cognitive diagnosis model is to mine learners' proficiency through the task of performance prediction. The proficiency vector is obtained by multiplying the learner's one-hot representation vector l with a trainable matrix $\boldsymbol{A}$. That is:

$$\boldsymbol{\theta} = \boldsymbol{l} \times \boldsymbol{A}. \tag{4}$$

3.2.3 Prediction of Learner Response

Interaction layer. The proposed method predicts a learner's response performance to a question as a probability. We input the representations of the learner parameter and question parameters (i.e., item discrimination and knowledge difficulty, respectively) into an interaction function to predict the learner's probability of answering the specific question correctly.

The interaction function simulates how learner parameters interact with question parameters to get the response results, for example, a simple logistic-like function is used as the interaction function in IRT. Based on previous works [22, 33, 34, 35], we use a neural network to learn a more complex non-linear interaction function to boost the model.

Specifically, the input of the interaction function can be formulated as:

$$\boldsymbol{x} = a\ (\boldsymbol{\theta} - \boldsymbol{b}) \odot \boldsymbol{k_c} \tag{5}$$

where $\boldsymbol{k_c}$ is the knowledge concept or skill vector that indicates the relationship between the question and knowledge concepts or skills, which is usually pre-labeled by experts and obtained directly from Q-matrix. We discuss how we define the knowledge concepts or skills in Section 3.3. The operator $\odot$ is the element-wise product and $\boldsymbol{x}$ indicates the learner's performance on each concept pertaining to the question. We then use a three-layer feed-forward neural network F_i to learn the non-linear activation function and output the probability p that the learner answers the question correctly. It can be formulated as:

$$p = \sigma(F_i(\boldsymbol{x})). \tag{6}$$

Following previous works [33, 34, 35], we restrict each weight of F_i to be positive during the process of training to ensure the monotonicity assumption, which assumes that the probability of learners answering the exercise correctly increases monotonically with the degree of mastery on each knowledge concept pertaining to the question.

Guess and Slip Adjustment. We noticed that many question items in the dataset are multiple-choice items, which makes it highly possible for the learners to guess the correct answer even if they don't master the knowledge concept, or slip even though they know the answer. To obtain better results, we add a guessing parameter $g \in [0,\ 1]$ and a slipping parameter $s \in [0,\ 1]$ to adjust the performance results, where g indicates the probability that a learner did not master the knowledge concepts but guessed the correct answer and s indicates the probability that a learner masters the knowledge concepts but did not answer correctly. The guessing and slipping parameters can be formulated as:

$$g = \sigma(F_g(\boldsymbol{i} \oplus \boldsymbol{l})), \tag{7}$$

$$s = \sigma(F_s(\boldsymbol{i} \oplus \boldsymbol{l})), \tag{8}$$

where F_g is the guessing and F_s is the slipping networks, respectively. To compute the final probability that a learner answers the question correctly, we apply adjustments of the guessing parameter and slipping parameter on the probability estimation, which can be expressed as:

$$y = g + (s - g) \times p. \tag{9}$$

3.2.4 Model Learning

We use the binary cross-entropy loss function for the proposed method. The learner's score is recorded as 1 when she/he answers the item correctly and 0 otherwise. For learner i and question j, let y_{ij} be the actual score for learner i on question j, and $\hat{y}_{ij}$ be the predicted score. Thus, the loss for learner i on question j is defined as:

$$\mathcal{L} = y_{ij} log\hat{y}_{ij} + (1 - y_{ij})log(1 - \hat{y}_{ij}). \tag{10}$$

Using Adam optimization [15], all parameters are learned simultaneously by directly minimizing the objective function. After training, the value of $\boldsymbol{\theta}$ is what we get as the diagnostic result, which denotes the learner's knowledge proficiency.

Table 1: An example subwords Q-matrix.

	Knowledge Concept									
Words	active	actual	actor	act	-tive	-tual	-ual	-tor	-or	⋯
active	1	0	0	1	1	0	0	0	0	⋯
actual	0	1	0	1	0	1	1	0	0	⋯
actor	0	0	1	1	0	0	0	1	1	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.3 Defining Knowledge Concepts and Skills

The knowledge concept or skill vector indicates the relationship between question items and knowledge concepts/skills, which is fundamentally essential as we need to diagnose the degree of proficiency of a learner corresponding to a specific knowledge concept/skill. As for each question, the knowledge concept/skill vector $\boldsymbol{c} = (c_1, c_2, c_3, \ldots c_k), c_i \in \{0, 1\}$ represents if a specific knowledge concept/skill is required to solve the question, in which $c_i = 1$ indicates that the knowledge concept/skill is included in the question and conversely $c_i = 0$ is not.

Usually, skills or knowledge concepts are pre-labeled by experts, and the vector $\mathbf{c}$ can be directly obtained from the pre-given Q-matrix. However, the knowledge concept/skill is difficult to define for language learning compared to other learning contexts such as science, engineering, and mathematics. Conventional models treat all question items nested under a particular word equivalent, but even for the same word, the ability of learners to comprehend a specific word can be divided into different levels. Some researchers define 'word knowledge' as different components including spelling, word parts, meaning, grammatical functions, the associations a word has with other words, and collocation to describe the totality of the learner's knowledge of a specific word in a language [20]. Thus, different items may refer to the same word if the word is used differently in multiple contexts (e.g., used as different parts of speech), or if different components of the word are tested. It is important to consider these when building vocabulary proficiency diagnosis models.

In the following subsections, we introduce several methods for defining knowledge concepts/skills in vocabulary proficiency diagnosis using different linguistic features and provide more detailed results on diagnosing associated knowledge concepts/skills.

3.3.1 Words as Knowledge Concepts

The simplest way to label knowledge concepts in an item is to simply use the unique words as knowledge concepts. There could be many knowledge concepts (e.g., many unique words) in a language-learning system, but only one knowledge concept (i.e., a word tested in the question) is related to a question item.

3.3.2 Sub-words as Knowledge Concepts

Another way to label multiple knowledge concepts in an item is to identify sub-words that comprise a word and treat each of these sub-words as an additional knowledge concept. Subwords can be viewed as morphological features of an original word, which may indicate the relationships of different

Table 2: Summary of question formats and required skill(s).

Format	Skill	Q-matrix Vector
F1	Recognition	[1, 0, 0, 0]
F2	Recognition	[1, 0, 0, 0]
F3	Recognition, Listening	[1, 1, 0, 0]
F4	Recognition, Spelling	[1, 0, 1, 0]
F5	Reading	[0, 0, 0, 1]

words and reinforce the knowledge related to gender agreement, prefixes, suffixes, compound words, etc. Inspired by the work of Zylich and Lan [38], we apply a sub-word tokenizer to automatically identify sub-words contained in each word. As shown in Table 1, we formulate a Q-matrix to apply the sub-word knowledge concepts for each word. For example, the word 'active' could have additional knowledge concepts such as 'act' and '-tive'.

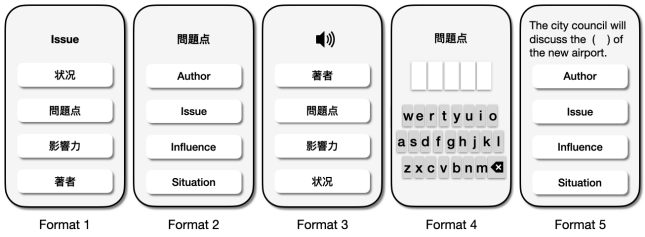

Figure 3: Examples of different question formats

3.3.3 *Semantically Similar Words as Knowledge Concepts*

Recent works indicated that cross-effects commonly exist in language learning [21, 38]. That is, during the exercise process of a learner, when an exercise of a particular knowledge concept is given, she/he also applies the relevant knowledge concepts to solve it. Specifically, in language learning, it seems that knowledge pertaining to semantically-similar words related to the word being tested are helpful in answering the question.

Following previous work [38], we used word embeddings to obtain semantic similarities of words. First, we embedded each word into a 300-dimensional vector using pre-trained fastText word embeddings [12] and calculated the cosine similarity scores between each pair of words to get a matrix of values that indicates the similarities of each word. Using this similarity matrix, all the similar words in the dataset that have cosine similarity larger than a threshold α with the current word can be counted as addition knowledge concepts required to solve the question. The threshold α is used to control the degree of semantic similarity, for example, only highly semantically similar words can be used as knowledge concepts in the Q-matrix if α is large, and if $\alpha = 1$, this model reduces back to the basic word-level model that only uses the current word as the knowledge concept. Otherwise, if $\alpha = 0$, which means that all other words that have non-negative similarity with the current word are treated as knowledge concepts.

3.3.4 *High-order Skills*

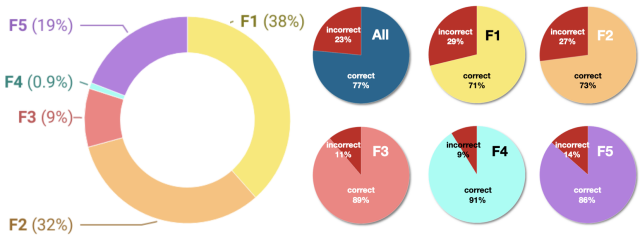

Figure 4: Distribution of question formats and response pie chart.

We formulated several methods for defining knowledge concepts in language proficiency diagnosis using different linguistic features such as additional morphological and semantic concepts. However, the ability used to solve vocabulary questions can depend on several high-order skills but not on whether the learner knows the word or not. Following previous works [14, 20, 37], we also consider defining skills instead of knowledge concepts in language proficiency diagnosis.

Here we propose two different methods to label skills in language proficiency. The most basic way we can choose to label a skill is by the question format. As shown in Figure 3, there are five different question formats in our dataset (more detailed information on the data can be found in Section 4.1). And if a learner is good at correctly answering a particular type of question, we can assume that she/he has a high skill in this question format. However, there will only be a single skill associated with each item and is not explainable enough if we use the question format as skills. To have a better interpretation, as summarized in Table 2, for each question format (see Figure 3), we defined some high-order language skills (i.e., Recognition, Listening, Spelling, Reading) required to tackle a specific question format based on some of the evidence from the literature [14, 16, 20, 26].

4. EVALUATION

4.1 Dataset

Our real-world dataset came from one of Japan's most popular English-language learning applications, and most of the users are Japanese students. The dataset includes 9,969,991 learner-item interactions from 2,014 users. There are 1,900 English words in the dataset, and each word has five different question formats collectively assessing different skills, resulting in 9500 items. The different question formats are shown in Figure 3, and some basic statistics of the dataset and response distributions are shown in Figure 4.

4.2 Experimental Settings

4.2.1 *Evaluation Metrics*

The performance of a cognitive diagnosis model is hard to evaluate as we can't obtain the true knowledge proficiency of learners directly. Usually, the models are evaluated by predicting learner performance in most cognitive diagnosis works. Following previous works, we evaluated by comparing the predicted responses with the ground truth, i.e., the actual response by the learners.

To set up the experiment, the data were randomly split into

80%/20% for training and test purposes, respectively. We filtered out the learners who had answered less than 50 questions so that every learner could be diagnosed with enough data. Like previous works [4, 34, 35], we use Prediction Accuracy (ACC), Area Under Curve (AUC), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) as metrics. The larger the values of ACC and AUC, and the smaller the values of MAE and RMSE, the better the results are.

4.2.2 Comparison

We name our model as Vocabulary Proficiency Diagnosis Model (VPDM) and compared our models using different knowledge concept and skill definitions with several existing models given below.

• DINA [6]: DINA is a cognitive diagnosis method that models learner concept proficiency by a binary vector.

• IRT [9]: IRT is a classical baseline method that models learners' and questions' parameters using the item response function.

• MIRT [24]: Extending from IRT, MIRT can model the multidimensional latent abilities of a learner.

• PMF [10]: Probabilistic matrix factorization (PMF) is a factorization method that can map learners and questions into the same latent factor space.

• NMF [17]: Non-negative matrix factorization (NMF) is also a factorization method, but it is non-negative, which can work as a topic model.

• NCD [33]: NCD is a recently proposed method that uses neural networks to learn more complex non-linear learner-question interaction functions.

Among these baselines, IRT, MIRT, and DINA are widely used methods in educational psychology. PMF and NMF are two matrix factorization methods from the recommendation system and data mining fields. NCD is a recently proposed model based on deep learning.

4.2.3 Parameter Settings

We implemented our model and other baselines in PyTorch. The model was trained with a batch size of 256. We used Adam optimizer with a learning rate of 0.001. The dropout rate is set to 0.2, and early stopping is applied to reduce overfitting.

5. RESULTS

5.1 Performance Prediction

The overall results on all four metrics are shown in Table 3 for all baseline methods and our models predicting learners' performance. VPDM-Word, VPDM-Subword, VPDM-Semantic, VPDM-FormatSkill, and VPDM-LangSkill are our models using words, subwords, semantically similar words, question formats, and language skills as knowledge concepts /skills, respectively. We observe that our models perform better than all other models, indicating the effectiveness of our framework. Among other baseline models, we noticed that the performance of NCD is comparable to our models and better than educational psychology methods (i.e., DINA, IRT, and MIRT) and matrix factorization methods (i.e., NMF and PMF), which demonstrates that leveraging deep learning could model the learner-question interactions more accurately than other conventional models.

Table 3: Performance comparison.

Model	ACC ↑	AUC ↑	MAE ↓	RMSE ↓
DINA	0.756	0.704	0.348	0.446
IRT	0.770	0.721	0.317	0.400
MIRT	0.768	0.728	0.311	0.399
NMF	0.768	0.722	0.355	0.405
PMF	0.771	0.731	0.328	0.398
NCD	0.772	0.734	0.316	0.397
VPDM-Word	0.773	0.736	0.309	0.396
VPDM-Subword	0.772	0.736	0.310	0.396
VPDM-Semantic	0.773	0.736	0.308	0.396
VPDM-FormatSkill	0.773	0.742	0.309	0.395
VPDM-LangSkill	0.773	0.742	0.308	0.395

In comparing our models, the performance of the VPDM-Word, VPDM-Subword, and VPDM-Semantic models are comparable, while VPDM-LangSkill and VPDM-FormatSkill models obtain better performance than other models, indicating that more broadly defined skills/knowledge concepts of an item are better. We will introduce our investigations to gain a deeper understanding of the difference among our models in the following subsections.

5.2 Impact of Different Formats

Many assessments have a mixture of item types (same as our dataset) since results based on a single format only reflect the knowledge unique to the specific format and might be misleading. To illustrate the performance of our models on different item formats, we separated the mixed-format dataset into different parts that only include different specific item formats, so we could conduct experiments to evaluate questions with a specific format. The results are shown in Figure 5 and the number of responses completed per learner is shown in Figure 6. Note that we did not test VPDM-LangSkill and VPDM-FormatSkill models here as they are intended for the mixed-format dataset.

Overall, the results indicate that our model consistently outperforms all other models. Furthermore, we observe that the prediction performance is affected by the question format, which highlights the fact that different question formats assess different traits.

5.3 Ablation Study

To investigate how the guessing and slipping adjustment layer affects model performance, we conducted some ablation experiments to compare the results. Table 4 shows the comparison results of the experiments on our mixed-format dataset and different single-format datasets. We observed that the performance improves when using the guessing parameter, and the model with guessing and slipping parameters obtained the best performance. It is reasonable as many items are multiple-choice in our dataset. In addition, we

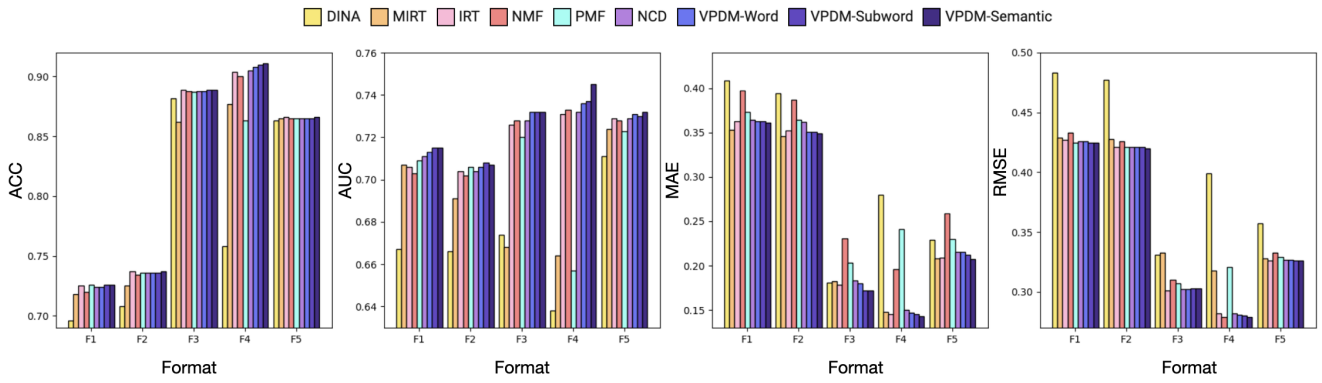

Figure 5: Comparison among different question formats.

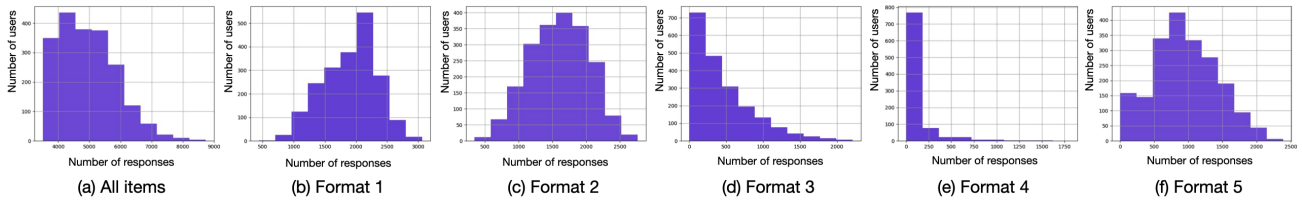

Figure 6: Distribution of the number of responses per learner.

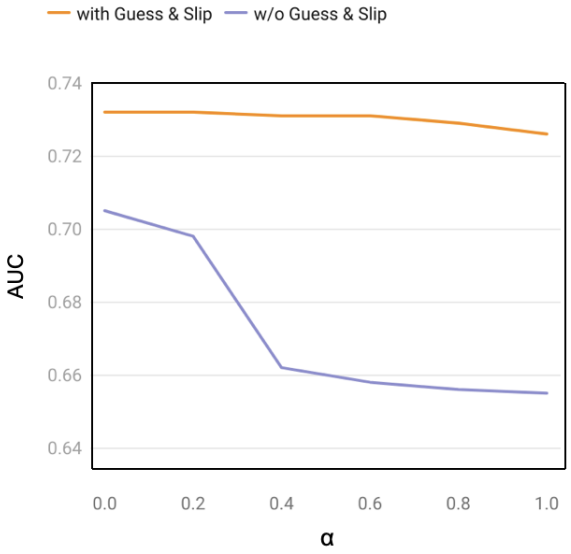

Figure 7: Comparative performance of semantically similar words as knowledge concepts via cosine similarity.

noticed that adding the slip and guessing parameters substantially improves some models' performance. This might imply that the Q-matrix is not specified appropriately in those models, though no formal rules exist to test this assumption [7].

In the comparison of the models that remove the guessing and slipping adjustment layer, the performance of the basic VPDM-Word model is the worst. As we expected, the knowledge assessed by a word item is not just simply related to the tested target word in the question. Moreover, the results confirm that the item's format carries meaning and is related to different traits, even though the questions with different formats are all designed for the same word.

As for subword and semantic models which use additional morphological or semantical knowledge concepts along with the tested target word, we observed improvements compared to the basic word-level model. One possible explanation is that the use of additional morphological or semantical knowledge concepts results in more items that share skills with each other, enabling the model to capture more interactions between learners and different words and reinforce the knowledge related to gender agreement, prefixes, suffixes, compound words, etc. [38]. For example, a closer inspection of the items revealed that even learners who are familiar with the word 'break' but do not know 'breakthrough' still have a good chance of answering some 'breakthrough' related items correctly. Figure 7 shows that varying the threshold parameter α in the VPDM-Semantic model does not influence the performance drastically. However, when we remove the guess and slip adjustment layer, we found that the performance of the model increases with the decreases of α, and the model performs best when $\alpha = 0$, which means that all other words that have non-negative similarity with the current word are treated as knowledge concepts. This result is in agreement with previous works, that an item designed to measure one trait may also require some level of other traits [37], and the proficiency of similar knowledge concepts can affect each other [11]. Specifically for language learning settings, it is important to focus not only on the interactions with the same word but also on interactions with other semantically similar words when predicting the degree of mastery of the target word [21]. We also noticed an intriguing finding for format 4, where VPDM-Subword and VPDM-Semantic outperformed the VPDM-Word model

Table 4: Results of the ablation study.

Model	Adjustment	All	F1	F2	F3	F4	F5
		AUC ↑	AUC ↑	AUC ↑	AUC ↑	AUC ↑	AUC ↑
VPDM-Word	-	0.655	0.668	0.669	0.685	0.628	0.715
	Guess	0.735	0.711	0.705	0.731	0.736	0.730
	Guess & Slip	0.736	0.713	0.706	0.732	0.736	0.731
VPDM-Subword	-	0.661	0.683	0.679	0.703	0.698	0.716
	Guess	0.734	0.711	0.703	0.729	0.731	0.729
	Guess & Slip	0.736	0.715	0.708	0.732	0.737	0.730
VPDM-Semantic	-	0.705	0.674	0.672	0.699	0.699	0.711
	Guess	0.734	0.713	0.706	0.730	0.732	0.731
	Guess & Slip	0.736	0.715	0.707	0.732	0.745	0.732
VPDM-FormatSkill	-	0.733					
	Guess	0.740	-	-	-	-	-
	Guess & Slip	0.742					
VPDM-LangSkill	-	0.735					
	Guess	0.741	-	-	-	-	-
	Guess & Slip	0.742					

significantly after the guess and slip adjustment layer was removed. This finding is particularly noteworthy because format 4 requires learners to type the word, and the results are more likely to be influenced by related morphological and semantic knowledge concepts such as prefixes, suffixes, and compound words. This result highlights the critical role of the item's format and how it influences the required knowledge in the question. Understanding this relationship between item format and knowledge requirements could potentially inform the design of more effective and efficient language learning assessments and improve learners' overall performance.

Finally, VPDM-LangSkill and VPDM-FormatSkill models obtain better performance than other models, indicating that more broadly defined skills and knowledge of an item are better in this task. For VPDM-FormatSkill model, one prevalent hypothesis is that items with different formats measure different traits or dimensions, and factors could be hypothesized to form on the basis of item format [31]. That is, the item's format might also be important and related to different traits or dimensions as suggested by previous works [7]. For VPDM-LangSkill model, the results show that learners' knowledge acquisition is influenced by high-order features (language abilities in this case). It greatly reduces the complexity of the model in cases where it is reasonable to view the examination as measuring several general abilities in addition to the specific knowledge states.

5.4 Interpretation of the Diagnosis

We visualize the diagnostic reports and evaluate the interpretation of the VPDM-LangSkill model as it is the most practical one with good performance. This visualization helps learners recognize their knowledge state intuitively and assists test developers to design question items effectively. As shown in Figure 8, we randomly sampled a learner and depict the proficiency diagnosed by IRT and VPDM-LangSkill. Each point on the radar diagram represents the mastery level of a certain trait. The red and blue lines denote the proficiency diagnosed by IRT and VPDM-LangSkill (scaled to $(0, 1)$), respectively. From the results, we can see that IRT only provides an overall unidimensional latent trait, the proficiency for all concepts is identical, therefore, it is not explainable enough to guide learners' self-assessment. As for the VPDM-LangSkill model, it is able to provide better interpretable insight for multidimensional traits (i.e., in our case, recognition, listening, spelling, and reading).

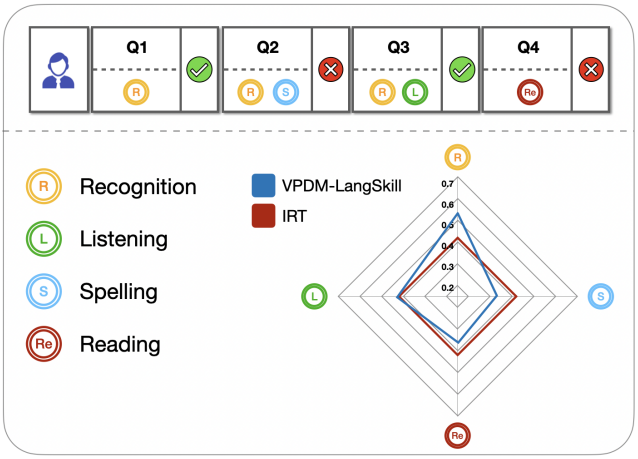

Figure 8: Visualization of a sample diagnostic report.

6. CONCLUSION

In this work, we proposed a framework for vocabulary proficiency diagnosis, which could capture the learner-question interactions more accurately using neural networks. In addition, we proposed a series of methods based on our framework, that uses different linguistic features to define skills and knowledge concepts in the context of a language learning task. Experimental results of cognitive diagnosis on real-

world second-language learning dataset showed that the proposed approach outperforms existing approaches with higher accuracy and increased interpretability. We also provided empirical evidence with ablation testing to prove that our knowledge concept and skill definitions are reasonable and critical to the performance of our model.

There are some limitations in this work. Firstly, the learner base of the dataset is limited to learners of the same language background and thus might decrease the generalize of this work. We plan to test other datasets in future work. In addition, we only consider the target word that is tested in the question, however, some questions are multiple-choice, and some questions test contextual usage as the learner needs to fill in a sentence with the correct target word. Therefore, additional features such as context information and distractors in the question should also be considered as they also influence the learner's performance. We expect that this work will provide useful implications for language-learning applications that focus on vocabulary learning, and we will test more question formats and include additional linguistic skills to expand the capabilities of our model in future work.

7. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP20H00622.

8. REFERENCES

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

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

[3] Y. Chen, Q. Liu, Z. Huang, L. Wu, E. Chen, R. Wu, Y. Su, and G. Hu. Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 989–998, 2017.

[4] S. Cheng, Q. Liu, E. Chen, Z. Huang, Z. Huang, Y. Chen, H. Ma, and G. Hu. 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, 2019.

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

[6] J. De La Torre. Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1):115–130, 2009.

[7] J. De La Torre and J. A. Douglas. Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3):333–353, 2004.

[8] M. C. Desmarais. Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, 13(2):30–36, 2012.

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

[10] N. Fusi, R. Sheth, and M. Elibol. Probabilistic matrix factorization for automated machine learning. *Advances in neural information processing systems*, 31, 2018.

[11] W. Gao, Q. Liu, Z. Huang, Y. Yin, H. Bi, M.-C. Wang, J. Ma, S. Wang, and Y. Su. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–510, 2021.

[12] É. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[13] Z. Huang, Q. Liu, Y. Chen, L. Wu, K. Xiao, E. Chen, H. Ma, and G. Hu. Learning or forgetting? a dynamic approach for tracking the knowledge proficiency of students. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–33, 2020.

[14] F. Kilickaya et al. Assessing l2 vocabulary through multiple-choice, matching, gap-fill, and word formation items. *Lublin Studies in Modern Languages and Literature*, 43(3):155–166, 2019.

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

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

[17] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

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

[19] Q. Liu, R. Wu, E. Chen, G. Xu, Y. Su, Z. Chen, and G. Hu. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(4):1–26, 2018.

[20] B. Ma, G. P. Hettiarachchi, and Y. Ando. Format-aware item response theory for predicting vocabulary proficiency. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 695–700, 2022.

[21] B. Ma, G. P. Hettiarachchi, S. Fukui, and Y. Ando. Each encounter counts: Modeling language learning and forgetting. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 79–88, 2023.

[22] H. Ma, J. Zhu, S. Yang, Q. Liu, H. Zhang, X. Zhang, Y. Cao, and X. Zhao. A prerequisite attention model for knowledge proficiency diagnosis of students. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4304–4308, 2022.

[23] A. Mnih and R. R. Salakhutdinov. Probabilistic matrix factorization. *Advances in neural information*

processing systems, 20, 2007.

[24] M. D. Reckase. Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer, 2009.

[25] F. Robertson. Word discriminations for vocabulary inventory prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1188–1195, 2021.

[26] L. S. Stæhr. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2):139–152, 2008.

[27] Y. Sun, S. Ye, S. Inoue, and Y. Sun. Alternating recursive method for q-matrix learning. In *Educational data mining 2014*, 2014.

[28] N. Thai-Nghe and L. Schmidt-Thieme. Multi-relational factorization models for student modeling in intelligent tutoring systems. In *2015 Seventh international conference on knowledge and systems engineering (KSE)*, pages 61–66. IEEE, 2015.

[29] S. Tong, Q. Liu, R. Yu, W. Huang, Z. Huang, Z. A. Pardos, and W. Jiang. Item response ranking for cognitive diagnosis.

[30] A. Toscher and M. Jahrer. Collaborative filtering applied to educational data mining. *KDD cup*, 2010.

[31] R. E. Traub. On the equivalence of the traits assessed by multiple-choice and constructed-response tests. *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, pages 29–44, 1993.

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

[33] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6153–6161, 2020.

[34] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Yin, S. Wang, and Y. Su. Neuralcd: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[35] X. Wang, C. Huang, J. Cai, and L. Chen. Using knowledge concept aggregation towards accurate cognitive diagnosis. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2010–2019, 2021.

[36] Z. Wang, Y. Gu, A. Lan, and R. Baraniuk. Varfa: A variational factor analysis framework for efficient bayesian learning analytics. *arXiv preprint arXiv:2005.13107*, 2020.

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

[38] B. Zylich and A. Lan. Linguistic skill modeling for second language acquisition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 141–150, 2021.