Investigating Concept Definition and Skill Modeling for Cognitive Diagnosis in Language Learning

Boxuan Ma
Kyushu University
Fukuoka, Japan
boxuan@artsci.kyushu-u.ac.jp

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

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

Shinichi Konomi
Kyushu University
Fukuoka, Japan
konomi@artsci.kyushu-u.ac.jp

Language proficiency diagnosis is essential to extract fine-grained information about the linguistic knowledge states and skill mastery levels of test takers based on their performance on language tests. Different from comprehensive standardized tests, many language learning apps often revolve around word-level questions. Therefore, knowledge concepts and linguistic skills are hard to define, and diagnosis must be well-designed. Traditional approaches are widely applied for modeling knowledge in science or mathematics, where skills or knowledge concepts are easy to associate with each item. However, only a few works focus on defining knowledge concepts and skills using linguistic characteristics for language knowledge proficiency diagnosis. In addressing this, we propose a framework for language 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 comprehensive experiments and analysis to prove that our knowledge concept and skill definitions are reasonable and critical to the performance of our model.

Keywords: cognitive diagnosis, language proficiency, linguistic skill, concept definition, skill modeling.

1. INTRODUCTION

Language proficiency diagnosis is one of the critical fundamental technologies supporting language education and has recently gained popularity in many language learning applications. Identifying the learners' latent proficiency level to higher accuracy is crucial in providing personalized materials and adaptive feedback (Avdiu et al., 2019). In practice, with the diagnostic results, systems can provide further support, such as learning planning, learning material recommendation, and computerized adaptive testing. Most importantly, it can help second-language

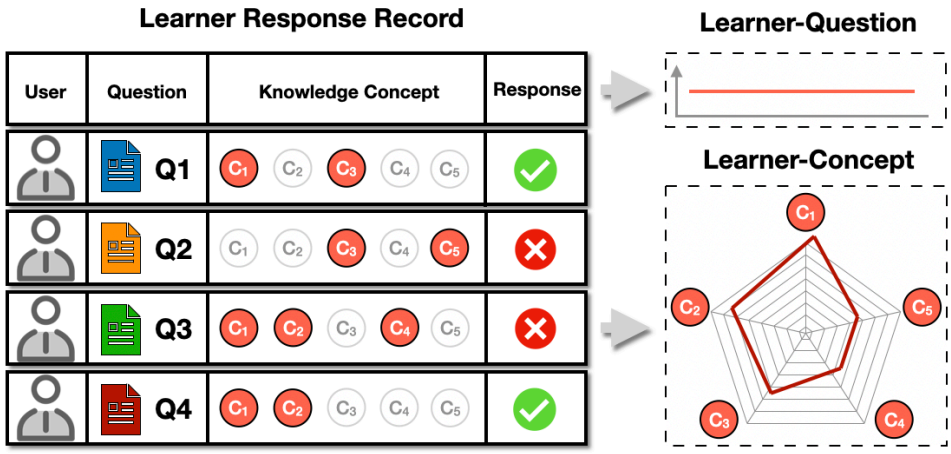

Figure 1: An example of cognitive diagnosis.

learners to place themselves in the correct learning space or level, especially when they are newcomers or have taken a long break from using the application, during which they might have forgotten a lot or, conversely, have advanced in the target language (Robertson, 2021).

Many cognitive diagnosis methods have been proposed for the 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 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 (Gao et al., 2021). For example, traditional methods like Item Response Theory (IRT) (Lord, 1980; Embretson and Reise, 2013), Multidimensional IRT (MIRT) (Reckase, 2009), and Matrix Factorization (MF) (Mnih and Salakhutdinov, 2007) try to model the learner-question interaction and provide learner latent traits (e.g., ability level) and the question features (e.g., difficulty level). Other works such as Deterministic Inputs, Noisy-And gate (DINA) (De La Torre, 2009) 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 by simply replacing the question with their corresponding concepts. Although great successes have been made, traditional methods have some limitations, which decay their effectiveness. Also, these approaches are widely applied for modeling knowledge in science or mathematics and ignore characteristics of language learning, which makes it a significant research challenge to infer the mastery level of learners' language 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 (Liu et al., 2019). For instance, IRT and MIRT focus only on analyzing learners' responses to determine the characteristics of

a question, such as its difficulty level, and the actual knowledge concepts being assessed in the question are not incorporated into these models. Most traditional methods were proposed for scale-based tests, where a group of examinees is tested using the same small set of questions. This ensures the completeness of response data, as each examinee is expected to answer every question. However, the data collection process for these traditional tests, such as paper-based exams, tends to be labor-intensive. Additionally, the diversity and breadth of question items are often constrained, given that all examinees are presented with the same examination paper. In contrast, 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 (Wang et al., 2020). Therefore, neglecting contents and formats leaves traditional methods no possibility to utilize the relationships of different items. Hence, they cannot generalize item parameters to unseen items (Robertson, 2021), while for language proficiency diagnosis, previous studies have shown that the information of questions is significantly related to item parameters, such as the difficulty level. For example, character length and corpus frequency are proven to be essential factors for predicting the difficulty level of questions (Culligan, 2015; Settles et al., 2020), while the average word and sentence lengths have been used as key features to predict text difficulty (Beinborn et al., 2014; Robertson, 2021). Also, studies have indicated that different question formats impact the difficulty level and explanatory power in predicting receptive skills (Kremmel and Schmitt, 2016). 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 essential to consider the format information of the items and their influence on different traits when building a language proficiency diagnosis model.

Another critical challenge is defining and using linguistic skills for diagnosing language proficiency. Although many approaches are widely applied for proficiency diagnosis, they have yet to be frequently 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 language learning is different from other domains since linguistic skills are hard to define (Ma et al., 2023a; Zylich and Lan, 2021). Unlike the comprehensive nature of standardized tests like TOEFL, many English learning apps often focus on simplistic word-level questions, which are easy to design and fast to answer by learners. However, defining knowledge concepts to extract intricate insights from seemingly basic word-level questions and providing detailed diagnoses remains a significant challenge.

To address these challenges, which have yet to be well explored in the research community, we propose a framework for language proficiency diagnosis, which could capture the learner-question interactions more accurately based on data mining and neural networks. In addition, we use linguistic features of words extracted by natural language processing tools, such as morphological and semantic features, to define knowledge concepts and skills related to vocabulary and grammar knowledge shared between words. Extensive experimental results on a real-world

second-language learning dataset demonstrate our proposed framework's effectiveness and interpretational power. 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 between different question formats and, in turn, their effect on vocabulary proficiency diagnosis. This paper is an extension of our previous conference paper (Ma et al., 2023b). Compared to our previous paper, we present more detailed experiment results in this paper. We also add a more comprehensive analysis and visualization of results to investigate the impact of question formats and the key factors that influence performance.

The remainder of this article is organized as follows. First, we will review the related work. Then, we introduce our task and present the framework in detail. Also, we introduce different ways to define knowledge concepts and skills. Next, we introduce the dataset and experimental settings in Section 4 and report the results in Section 5. Finally, we discuss the results and the limitations of our work in Section 6 and conclude the paper in Section 7.

2. Related work

2.1. Cognitive diagnosis

Cognitive diagnosis is a fundamental task, and many classical cognitive diagnosis models have been developed in educational psychology, such as IRT, MIRT, and DINA. IRT (Lord, 1980; Embretson and Reise, 2013) is a widely used method and has been applied in educational testing environments since the 1950s (Embretson and Reise, 2013). 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, it only provides an overall latent trait for learners, while each question usually assesses different knowledge concepts or skills, and a single ability dimension is sometimes insufficient to capture the relevant variation in human responses. By extending the trait features into multidimensions, Reckase (2009) proposed MIRT, which tries to meet multidimensional data demands by including an individual's multidimensional latent abilities for each skill. Although IRT and MIRT are powerful for assessing learner abilities based on their responses, the latent trait vectors they provide are often abstract and not directly interpretable in a practical educational context. For instance, a high latent trait score in a mathematics assessment might indicate a general mathematical ability, but it doesn't specify areas of strength or weakness in a way that is actionable for a learner's self-assessment (Cheng et al., 2019; Wang et al., 2023).

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 (Ma et al., 2022). To provide detailed results on each knowledge concept or skill, other

works try to build learner-concept interaction directly. For example, DINA (De La Torre, 2009) models 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. Unlike IRT, θ and β are multidimensional and binary in DINA, where β came directly from Q-matrix. Two other parameters, guessing g and slipping s, are also considered. The guessing parameter reflects the likelihood that a student will correctly answer an item by guessing despite not having the required skills or knowledge. In contrast to guessing, the slipping parameter represents the probability that a student with the required skills or knowledge for a particular item will answer incorrectly. This could happen due to various reasons, such as making careless mistakes, misinterpreting the question, or temporary confusion. The DINA formula is written as follows:

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

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 or does not master all the required skills but makes a successful guess. Although DINA has made significant 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 recommendation systems (Chen et al., 2017). 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 and Jahrer (2010) 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, Thai-Nghe and Schmidt-Thieme (2015) proposed multi-relational factorization models to exploit multiple data relationships to improve the prediction results in intelligent tutoring systems. In addition, Desmarais (2012) used Non-negative Matrix Factorization (NMF) to map question items to skills, and the resulting factorization allows a straightforward interpretation of a Q-matrix. Similarly, Sun et al. (2014) proposed a method that uses Boolean Matrix Factorization (BMF) to map items into latent skills based on learners' responses. Wang et al. (2020) 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 (Huang et al., 2020; Liu et al., 2018; Tong et al., 2021; Tong et al., 2022).

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. (2020) proposes a Neural Cognitive Diagnosis (NCD) framework for Intelligent Education Systems, which leverages neural networks to learn the interaction function automatically.

Some researchers focus on incorporating the content representation from question texts into the model by neural networks, which is difficult with traditional methods. Cheng et al. (2019) 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. (2023) applied neural networks to extract two typical types of information in the question text: knowledge concepts and extra text-related factors. Furthermore, Song et al. (2023) utilized semantic information in the cross-modal contents of exercises for modeling student performance. 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 to enhance diagnosis performance. For example, Wang et al. (2021) proposed a model based on neural networks and aggregate knowledge relationships by converting all knowledge concepts into a graph structure. Ma et al. (2022) proposed the Prerequisite Attention model for Knowledge Proficiency (PAKP) to explore the prerequisite relation among knowledge concepts and use it for inferring knowledge proficiency. Li et al. (2022) proposed a Bayesian network-based Hierarchical Cognitive Diagnosis Framework (HierCDF) to incorporate knowledge attribute hierarchy when assessing students. Recent work proposed the Relation map driven Cognitive Diagnosis (RCD) (Gao et al., 2021) model by comprehensively modeling the learner-question interactions and question-concept relations. Their model performed better than traditional works considering only learner-question interactions (e.g., IRT) or only question-concept interactions (e.g., DINA).

When looking into studies about language proficiency assessments, the application of Natural Language Processing (NLP) has recently become increasingly prevalent. Techniques such as word embedding (e.g., word2vec), Recurrent Neural Networks (RNNs), and Bidirectional Encoder Representations from Transformers (BERT) have been instrumental in extracting and interpreting semantic information from texts, and they are essential for applications of assessing language abilities. For example, studies have investigated the relationship between test items and language features to predict the item's difficulty or even automatically generate new items. Susanti et al. (2016) proposed a system to generate questions pertaining to vocabulary. Factors such as the reading passage difficulty, semantic similarity between the correct answer and distractors, and distractor word difficulty level are all considered in this system. Beinborn et al. (2014) used NLP techniques to predict c-test difficulty at the word-gap level, combining phonetic and text complexity factors. Loukina et al. (2016) 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. (2020) used a Markov chain language model and unigram language model

to induce linguistic features and then used machine learning models to estimate item difficulty directly. Benedetto et al. (2021) used pre-trained BERT models to estimate the difficulty of multiple-choice language questions. Other studies try to estimate learners' knowledge states. For example, Ma et al. (2023a) models and predicts learners' knowledge by considering their forgetting behavior and linguistic features in language learning. Pre-trained word embeddings are used in their model to extract semantic and morphological features. There is an increasing amount of work that uses machine learning and NLP approaches for language proficiency assessments. However, most of these studies focus on generating question items or predicting learners' responses using linguistic features extracted by NLP methods, and few directly use linguistic features to define knowledge concepts and skills. Also, these studies failed to consider question formats' influence on different receptive skills.

In summary, 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 for language proficiency diagnosis. In addition, except for the work by Wang et al. (2023), other works mentioned above failed to consider question formats, which are essential for language-learning questions and may significantly influence 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 is 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, three parts need to be considered: learner, question item, and interaction function. As shown in Figure 2, we propose a cognitive diagnostic framework with deep learning 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, question, and the knowledge concepts vector of the question as input. Then, we obtain the diagnostic parameters of the learner and question. Next, 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.

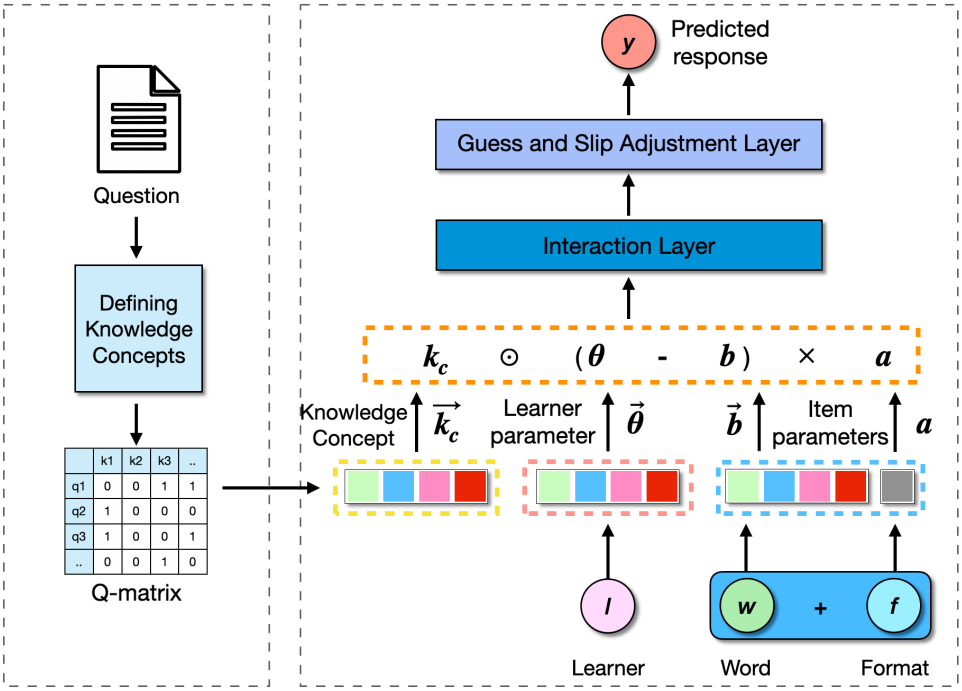

Figure 2: Overview of the proposed framework.

3.2.1. Item parameters

The item's characteristics are calculated in the item network to represent the traits of a specific item. Our model uses two parameters from the Two-Parameter Logistic IRT model (Van der Linden and Hambleton, 1997), 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 mentioned, 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 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, IRT uses a simple logistic-like function as the interaction function. Based on previous works (Ma et al., 2022; Wang et al., 2020; Wang et al., 2023; Wang et al., 2021), 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 follows:

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

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 about 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 follows:

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

Following previous works (Wang et al., 2020; Wang et al., 2023; Wang et al., 2021), 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 about the question. Next, we used the sigmoid function for output binary learner responses (e.g., 0 or 1).

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 do not 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 follows:

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

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

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 scores 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 (Kingma and Ba, 2014), 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.

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, $c_i = 0$ indicates it is not included.

Usually, skills or knowledge concepts are pre-labeled by experts, and the vector $\boldsymbol{c}$ 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 as equivalent. However, 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 (Nation, 2001; Ma et al., 2023a). 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 essential to consider these when building vocabulary proficiency diagnosis models.

The following subsections introduce several methods for defining knowledge concepts/skills in vocabulary proficiency diagnosis using different linguistic features. We also provide more detailed results on diagnosing associated knowledge concepts/skills.

Table 1: An example subwords Q-matrix.

Words	**Knowledge Concept**									
	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	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.3.1. Words as knowledge concepts

The simplest way to label knowledge concepts in an item is to use the unique words as lexical knowledge concepts. A language-learning system could have many lexical knowledge concepts (e.g., many unique words), and actually, many questions are designed based on a word. Therefore, we use the word that is tested in the question as the knowledge concept, and only one knowledge concept will relate to a question item of this type.

3.3.2. Sub-words as knowledge concepts

In reality, multiple knowledge concepts are often required to successfully solve a problem, and using the tested word as the only knowledge concept for a question item may not be enough. For example, to achieve proficiency in a compound word, it often necessitates mastery of its constituent sub-words. In order to label multiple knowledge concepts in an item, we identify sub-words that comprise a word and treat each sub-word as an additional knowledge concept. Sub-words can be viewed as morphological features of an original word, which may indicate the relationships of different words and reinforce the knowledge related to gender agreement, prefixes, suffixes, compound words, etc. Inspired by the work of Zylich and Lan (Zylich and Lan, 2021), we apply a sub-word tokenizer to identify sub-words contained in each word automatically. 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 'act' and '-tive'.

3.3.3. Semantically similar words as knowledge concepts

Previous works indicated that cross-effects commonly exist in language learning (Ma et al., 2023a; Zylich and Lan, 2021). 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 is helpful in answering the question.

To obtain semantic similarities of words, we first embedded each word into a 300-dimensional vector using pre-trained fastText word embeddings (Grave et al., 2018) and calculated the cosine similarity scores between each pair of words to get a matrix of values that indicate 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 additional 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

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

Format&Skill		**Q-matrix Vector**			
Format	Skill	Q1 Recognition	Q2 Listening	Q3 Spelling	Q4 Reading
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

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 $\alpha = 0$, all other words with non-negative similarity to the current word are treated as knowledge concepts.

3.3.4. High-order skills

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 to solve vocabulary questions can sometimes depend on several high-order skills rather than on whether the learner knows the word. Following previous works (Kilickaya et al., 2019; Ma et al., 2022; Yao and Schwarz, 2006), 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. Figure 3 shows five different question formats in our dataset (more detailed information on the data can be found in Section 4.1). Moreover, if a learner correctly answers 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 it 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 (Kilickaya et al., 2019; Kremmel and Schmitt, 2016; Ma et al., 2022; Stæhr, 2008).

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, and each row includes a user id, item id, question format, and response result (0 or 1). 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.

As shown in Figure 3, there are five different question formats to collectively assess reading, writing, listening, and speaking skills of vocabulary learning in our dataset. Below are the descriptions of the five question formats. Format 1: Multiple-choice, choose the correct Japanese

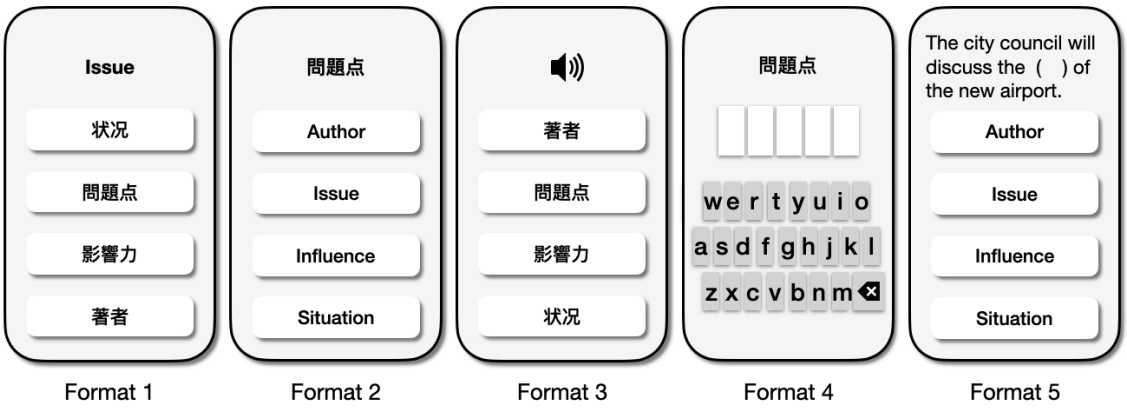

Figure 3: Examples of different question formats.

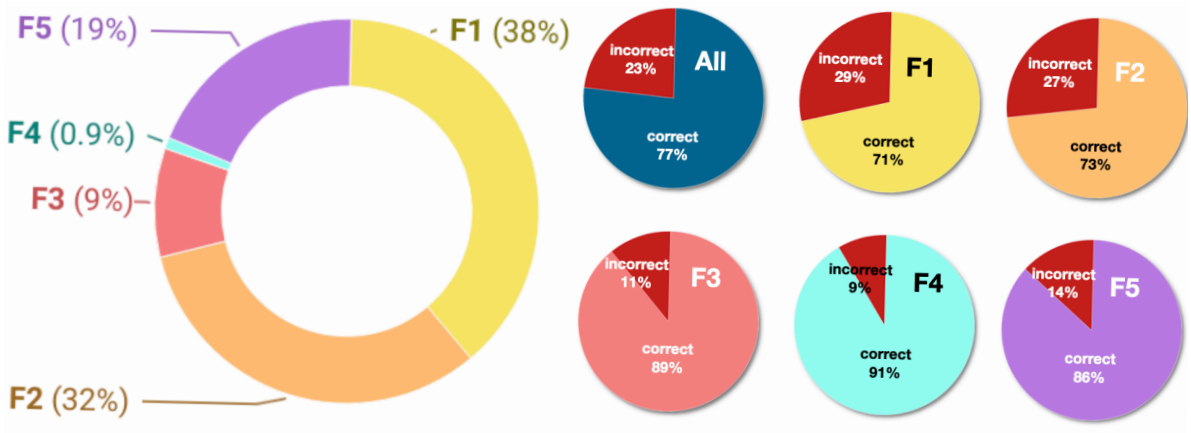

Figure 4: Distribution of question formats and response pie chart.

description of the English word. Format 2: Multiple-choice, choose the corresponding English word given the Japanese description. Format 3: Multiple-choice, listen to an English word, and choose the corresponding Japanese description. Format 4: Typing, type in letters to form the correct English word given the Japanese description. Format 5: Multiple-choice, choose the appropriate English word to fill in the blank of the sentence.

Moreover, some basic statistics of the dataset and response distributions are depicted in Figure 4. We notice an imbalance in the response data for different question formats, primarily influenced by the learning content designs, such as lessons and practice sessions within the application. Formats 1 and 2 are the most common in the datasets, accounting for 38% and 32%, respectively. Additionally, while the correctness rates are high across all question formats, there is a noticeable variance. The average correctness rate for the dataset is 77%, with format 4 achieving the highest average correctness rate of 91% and format 1 presenting the lowest at 71%. This variation indicates differing levels of difficulty and learner engagement with each question format.

4.2. EXPERIMENTAL SETTINGS

4.2.1. Evaluation metrics

The performance of a cognitive diagnosis model is hard to evaluate as we cannot 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 (Cheng et al., 2019; Wang et al., 2023; Wang et al., 2021), 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 named our model as Vocabulary Proficiency Diagnosis Model (VPDM) and compared our models using different knowledge concepts and skill definitions with several existing models given below.

- DINA (De La Torre, 2009): DINA is a cognitive diagnosis method that models learner concept proficiency by a binary vector.
- IRT (Embretson and Reise, 2013): IRT is a classical baseline method that models learners' and questions' parameters using the item response function.
- MIRT (Reckase, 2009): Extending from IRT, MIRT can model the multidimensional latent abilities of a learner.
- PMF (Fusi et al., 2018): Probabilistic matrix factorization (PMF) is a factorization method that can map learners and questions into the same latent factor space.
- NMF (Lee and Seung, 2000): Non-negative matrix factorization (NMF) is also a factorization method, but it is non-negative, which can work as a topic model.
- NCD (Wang et al., 2020): 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 [1]. 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.

[1]The code is available at https://github.com/BoxuanMa/Vocabulary-Proficiency-Diagnosis-Model

Table 3: Performance comparison. In the table, an upward arrow indicates that a higher value is better. Conversely, a downward arrow signifies that a lower value is better.

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

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, sub-words, 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.

In comparing our models, the performance of the VPDM-Word, VPDM-Subword, and VPDM-Semantic models are comparable. In contrast, VPDM-LangSkill and VPDM-FormatSkill models perform better than others, indicating that more broadly defined skills/knowledge concepts of an item are better. In the following subsections, we will introduce our investigations to gain a deeper understanding of the differences among our models.

5.2. 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. 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 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 (De La Torre and Douglas, 2004).

Table 4: Results of the ablation study. In the table, an upward arrow indicates that a higher value is better. Conversely, a downward arrow signifies that a lower value is better.

Model	Adjustment	ACC ↑	AUC ↑	MAE ↓	RMSE ↓
Word	-	0.765	0.655	0.343	0.412
	G	0.771	0.735	0.311	0.397
	G&S	0.773	0.736	0.309	0.396
Subword	-	0.766	0.661	0.343	0.412
	G	0.772	0.734	0.317	0.397
	G&S	0.772	0.736	0.316	0.396
Semantic	-	0.766	0.705	0.327	0.404
	G	0.772	0.734	0.310	0.397
	G&S	0.773	0.736	0.308	0.396
Format	-	0.772	0.733	0.319	0.399
	G	0.773	0.740	0.312	0.395
	G&S	0.773	0.742	0.309	0.395
LangSkill	-	0.770	0.735	0.315	0.397
	G	0.773	0.741	0.311	0.395
	G&S	0.773	0.742	0.308	0.395

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 sub-word and semantic models, which use additional morphological or semantic 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 semantic 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. (Zylich and Lan, 2021). For example, a closer inspection of the items revealed that even learners who are

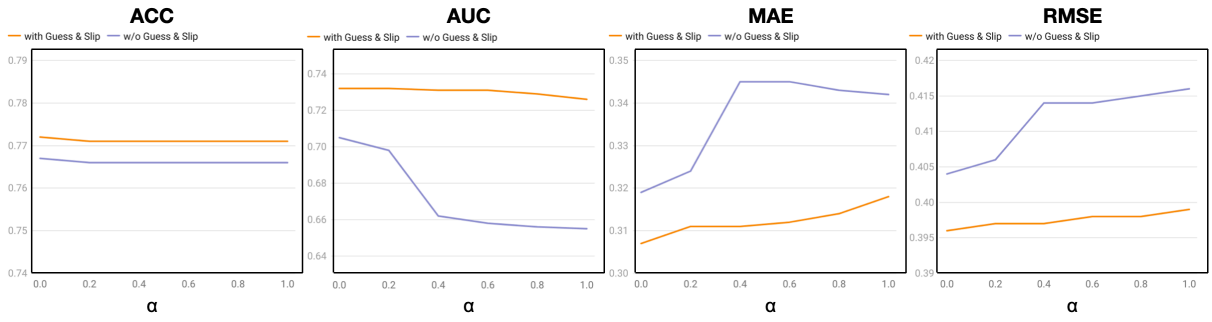

Figure 5: Comparative performance of semantically similar words as knowledge concepts via cosine similarity.

Table 5: Results of the ablation study for different question formats.

Format		Model											
		Word				Subword				Semantic			
		ACC	AUC	MAE	RMSE	ACC	AUC	MAE	RMSE	ACC	AUC	MAE	RMSE
F1	-	0.718	0.668	0.374	0.436	0.718	0.683	0.378	0.433	0.718	0.674	0.373	0.435
	G	0.723	0.711	0.364	0.426	0.723	0.711	0.365	0.427	0.724	0.713	0.364	0.426
	G&S	0.725	0.713	0.363	0.426	0.726	0.715	0.363	0.425	0.726	0.715	0.361	0.425
F2	-	0.731	0.669	0.369	0.428	0.734	0.679	0.364	0.426	0.733	0.672	0.364	0.428
	G	0.735	0.705	0.350	0.421	0.735	0.703	0.356	0.422	0.737	0.706	0.353	0.421
	G&S	0.737	0.706	0.351	0.421	0.736	0.708	0.351	0.421	0.737	0.707	0.349	0.420
F3	-	0.888	0.685	0.188	0.308	0.888	0.703	0.164	0.308	0.888	0.699	0.164	0.308
	G	0.888	0.731	0.180	0.302	0.889	0.729	0.190	0.303	0.889	0.730	0.187	0.304
	G&S	0.888	0.732	0.180	0.302	0.889	0.732	0.172	0.303	0.889	0.732	0.172	0.303
F4	-	0.907	0.628	0.146	0.290	0.907	0.698	0.142	0.287	0.907	0.699	0.142	0.284
	G	0.907	0.736	0.146	0.281	0.907	0.731	0.143	0.282	0.907	0.732	0.143	0.284
	G&S	0.908	0.736	0.147	0.281	0.907	0.737	0.150	0.280	0.908	0.745	0.143	0.279
F5	-	0.866	0.715	0.217	0.329	0.866	0.716	0.220	0.328	0.866	0.711	0.222	0.329
	G	0.866	0.730	0.216	0.327	0.866	0.729	0.213	0.326	0.866	0.731	0.213	0.326
	G&S	0.866	0.731	0.215	0.327	0.866	0.731	0.212	0.326	0.866	0.732	0.207	0.326

familiar with the word 'break' but do not know 'breakthrough' still have a good chance of answering some 'breakthrough' related items correctly. Figure 5 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 (Yao and Schwarz, 2006), and the proficiency of similar knowledge concepts can affect each other (Gao et al., 2021). 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 (Ma et al., 2023a).

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 (Traub, 1993). That is, the item's format might also be meaningful and related to different traits or dimensions as suggested by previous works (De La Torre and Douglas, 2004). 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.

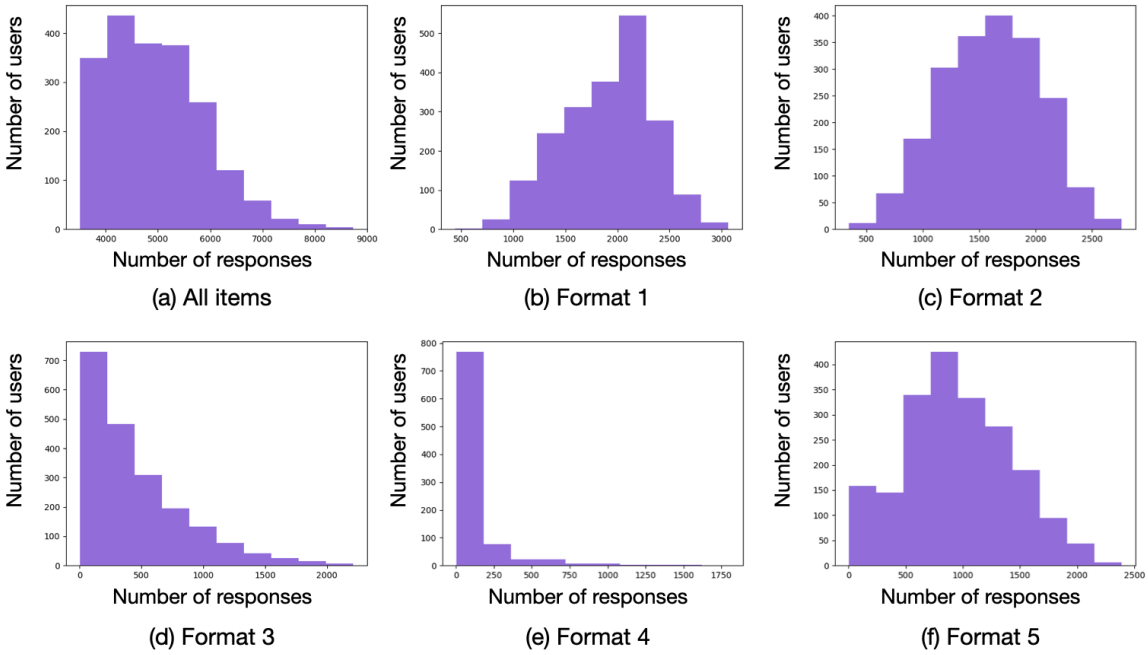

Figure 6: Distribution of the number of responses per learner.

5.3. 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 number of responses completed per learner is shown in Figure 6, and the comparison results are shown in Figure 7. 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

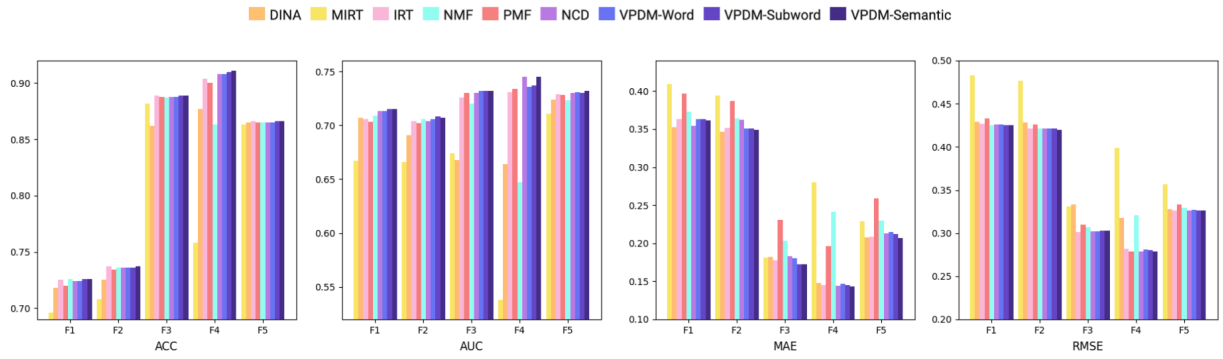

Figure 7: Comparison among different question formats.

outperforms all other models for different single-format datasets. Furthermore, we observe that the prediction performance is affected by the question format, which highlights the fact that different question formats assess different traits.

We also conducted ablation experiments to compare the results for different single-format datasets. Table 5 shows the comparison results of the experiments. Similar to the results for the mixed-format dataset, we noticed performance improvement when using the guessing and slipping parameters. The model that includes both guessing and slipping parameters achieved the best performance for different formats. Among models removing the guessing and slipping adjustment layer, the VPDM-Word model showed the worst performance. VPDM-Subword and VPDM-Semantic models, utilizing added morphological or semantic knowledge concepts alongside the target word, outperformed the basic word-level model.

In addition, we noticed intriguing findings across different formats. For example, VPDM-Subword and VPDM-Semantic outperformed the VPDM-Word model significantly after the guess and slip adjustment layer was removed for format 4, compared to other formats. 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. Conversely, for format 5, the performance of VPDM-Subword and VPDM-Semantic are almost the same as the VPDM-Word model, with or without the guess and slip adjustment layer. The possible explanation is, format 5 assesses broader comprehensive reading skills, requires learners to choose the correct word to fill the blank of a sentence, and does not directly test the target word itself. Therefore, the morphological and semantic knowledge concepts related to the target word may not be substantial enough to help learners understand the meaning of the entire sentence and correctly answer the question. That is, if learners fail to grasp the meaning of the complete sentence, they may not be able to provide the correct response, even if they possess knowledge of the target word.

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.

5.4. Parameter analysis

After training the model, we visualize the item and learner parameters to understand our models better. For multi-dimensional parameters, t-SNE (Van der Maaten and Hinton, 2008) is used to project them to 2-D points.

5.4.1. Item parameters

Figure 8 shows the difficulty visualization result of VPDM-Word, VPDM-LangSkill, and VPDM-FormatSkill models, colored with each vector by the corresponding question format. The results show that the distribution of the difficulty parameter vector is closely related to the corresponding question format for each model, even for the basic VPDM-Word model. As for VPDM-LangSkill and VPDM-FormatSkill models, since the high-order skills are defined by question formats, we can see that questions are clearly clustered by corresponding formats.

We also present the discrimination, guess, and slip parameters, as shown in Figure 9. We can observe the obvious differences between question formats. Generally, the discrimination parameters of format 5 are larger than others as it assesses broader comprehensive skills than

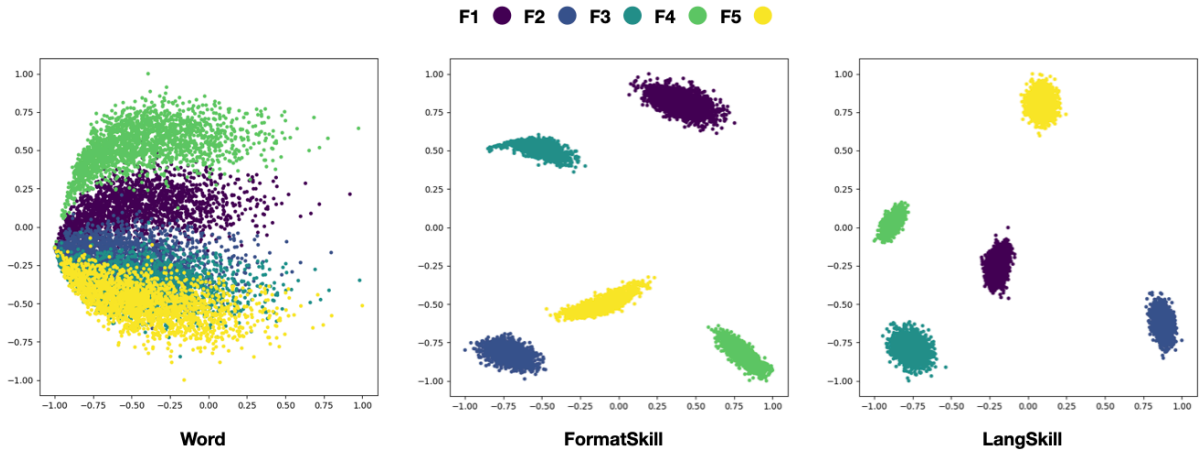

Figure 8: Difficulty visualization result of VPDM-Word model, VPDM-FormatSkill model, and VPDM-LangSkill model.

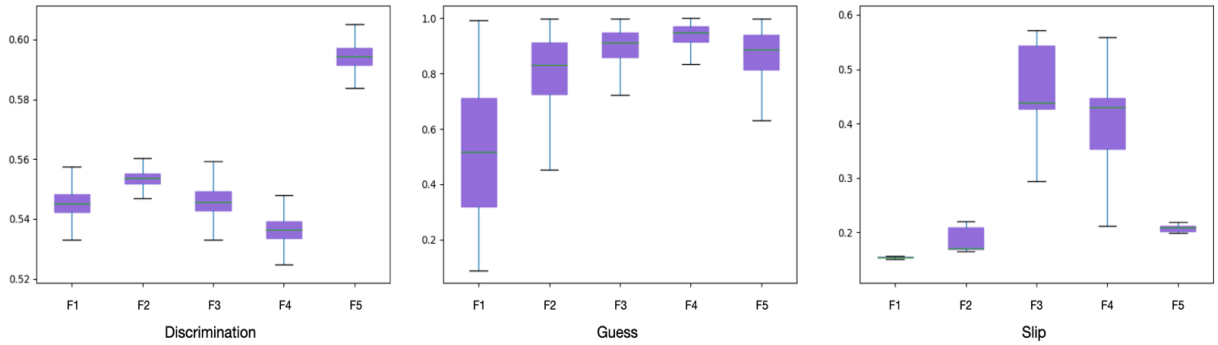

Figure 9: Distribution of discrimination, guess, and slip parameters.

other formats. The guess parameters are high across all formats due to the multiple-choice nature of the questions. Moreover, slip parameters are high for format 3 and format 4. These visualized results are consistent with the intuitive fact that it is much easier to slip for typing word questions.

5.4.2. Learner parameter

To further observe the relationship between learner parameter and performance, we visualize the learner parameter vectors of VPDM-LangSkill model and VPDM-FormatSkill model using t-SNE and color each vector by the corresponding learners' average response score in Figure 10. The distribution of learner parameter vectors for two models is closely related to their average scores, where we could observe that the points follow specific patterns instead of scattering randomly. The average scores gradually increase from right to left for both two models.

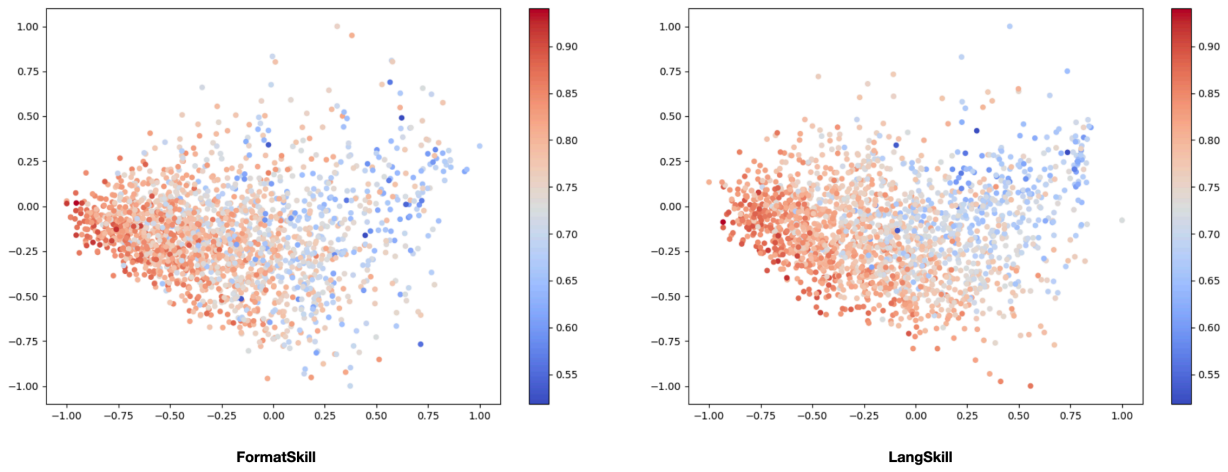

Figure 10: Proficiency visualization by the average score.

5.5. Interpretation of the Diagnosis

We visualized the diagnostic reports and evaluated 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 in designing question items effectively. As shown in Figure 11, we randomly sampled a learner and depicted the proficiency diagnosed by IRT and VPDM-LangSkill. We selected IRT here because it is a widely used method and has been implemented in the application. Specifically, we use a radar chart to show the proficiency of a learner on concepts diagnosed by IRT, and VPDM-LangSkill. Each point on the radar diagram represents the mastery level of a specific 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, and the proficiency for all concepts is identical. Therefore, it is not explainable enough to guide learners' self-assessments. As for the VPDM-LangSkill model, it can provide better interpretable insight for multidimensional traits (i.e., in our case, recognition, listening, spelling, and reading).

6. Discussion

Language proficiency diagnosis plays an important role in the field of language learning, which aims to identify the level of 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. Distinguishing from the comprehensive nature of standardized assessments like TOEFL, many English learning applications only provide word-level questions. Despite their apparent simplicity, these questions present a challenge in terms of defining detailed knowledge concepts and providing comprehensive diagnoses. In contrast to fields such as science or mathematics, where skills or knowledge concepts are well-defined and easy to associate with each item, the task of associating linguistic characteristics with skills and concepts for language knowledge proficiency diagnosis using word-level questions still needs to be explored. To tackle this issue, we propose a novel framework for language proficiency diagnosis based on neural

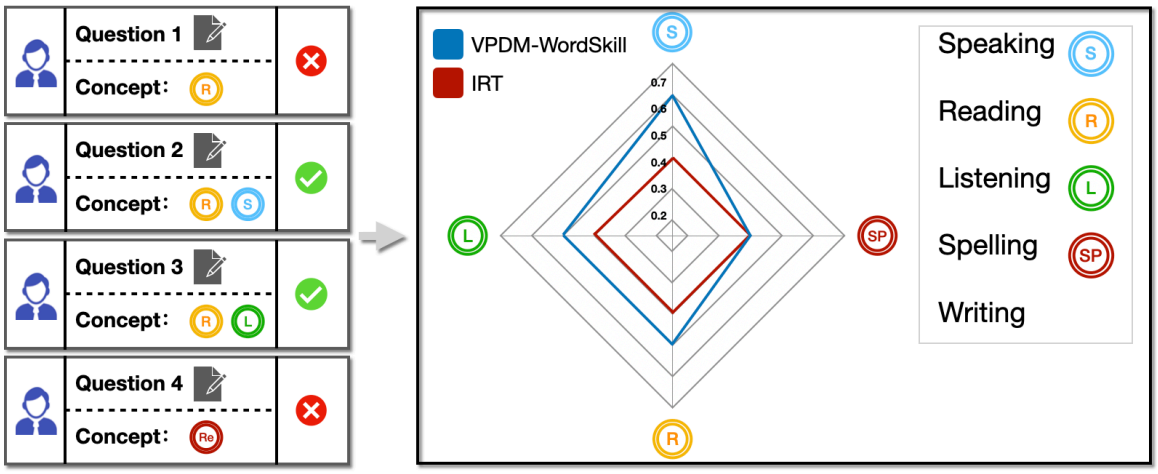

Figure 11: Visualization of a sample diagnostic report.

networks. Within this framework, we propose a series of methods based on our framework that use 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 comprehensive experiments and analysis to prove that our knowledge concept and skill definitions are reasonable and critical to the performance of our model.

Our main findings are as follows. First, we formulated several methods for defining knowledge concepts in language proficiency diagnosis using different linguistic features, such as additional morphological and semantic concepts. The results show that incorporating additional morphological and semantic features of each word improves the models. As expected, the knowledge assessed by a word item is not just related to the tested target word in the question. Second, we explored various approaches to provide broadly defined skills instead of knowledge concepts in language proficiency diagnosis. These models perform better than other specifically defined linguistic feature-level knowledge concept models. This result indicates that more broadly defined skills of an item are better in this task as high-order language abilities influence learners' knowledge acquisition. It is reasonable to view the language examination as measuring several general abilities in addition to the specific knowledge states, even for the word-level questions. Furthermore, 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. 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.

Our results can provide helpful data-driven insights into better language learning experiences. Concept definition and skill modeling for cognitive diagnosis in language learning explored in this paper can benefit question designers and learners.

6.1. LIMITATION AND FUTURE WORK

There are some limitations in this work. Firstly, the learner base of the dataset is limited to learners of the same language background, and the response data for different question formats is imbalanced. This might decrease the generalization of this work. We plan to test other datasets in future work. In addition, we only consider the target word 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 influence the learner's performance. Also, it is likely that the five 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 question formats and include additional linguistic skills to expand the capabilities of our model in future work. Finally, with a deeper observation of the results, we noticed that sometimes different models provide distinct parameters and diagnostic results as knowledge concepts are defined differently. This raises the question of the reliability of these results and how to appropriately apply the diagnosed outcomes. As our models are data-driven, proficiencies diagnosed by different models are not strictly guaranteed to be comparable (Wang et al., 2023). The explanation and usage of diagnosed proficiencies should be further explored. We leave the comparison of diagnosed proficiencies from different trained models and the validation of their credibility for future exploration.

7. CONCLUSION

In this work, we proposed a framework for language proficiency diagnosis, which could capture the learner-question interactions more accurately using neural networks. Within this framework, we proposed a series of methods based on our framework that incorporates different linguistic features to define skills and knowledge concepts for each word 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 and parameter analysis to prove that our knowledge concept and skill definitions are reasonable and critical to the performance of our model. We expect this work will provide valuable implications for language-learning applications.

8. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP20H00622, and the OpenDNA joint research program.

REFERENCES

AVDIU, D., BUI, V., AND KLIMČÍKOVÁ, K. P. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, D. Alfter, E. Volodina, L. Borin, I. Pilan, and H. Lange, Eds. LiU Electronic Press, 1–9.

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.

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*, J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, Eds. Association for Computational Linguistics, 147–157.

CHEN, Y., LIU, Q., HUANG, Z., WU, L., CHEN, E., WU, R., SU, Y., AND HU, G. 2017. Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Association for Computing Machinery, 989–998.

CHENG, S., LIU, Q., CHEN, E., HUANG, Z., HUANG, Z., CHEN, Y., MA, H., 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*. Association for Computing Machinery, 2397–2400.

CULLIGAN, B. 2015. A comparison of three test formats to assess word difficulty. *Language Testing 32,* 4, 503–520.

DE LA TORRE, J. 2009. Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics 34,* 1, 115–130.

DE LA TORRE, J. AND DOUGLAS, J. A. 2004. Higher-order latent trait models for cognitive diagnosis. *Psychometrika 69,* 3, 333–353.

DESMARAIS, M. C. 2012. Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter 13,* 2, 30–36.

EMBRETSON, S. E. AND REISE, S. P. 2013. *Item response theory*. Psychology Press.

FUSI, N., SHETH, R., AND ELIBOL, M. 2018. Probabilistic matrix factorization for automated machine learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, and N. Cesa-Bianchi, Eds. Curran Associates Inc., 3352–3361.

GAO, W., LIU, Q., HUANG, Z., YIN, Y., BI, H., WANG, M.-C., MA, J., WANG, S., AND SU, Y. 2021. 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*. Association for Computing Machinery, 501–510.

GRAVE, É., BOJANOWSKI, P., GUPTA, P., JOULIN, A., AND MIKOLOV, T. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. European Language Resources Association (ELRA), 3483–3487.

HUANG, Z., LIU, Q., CHEN, Y., WU, L., XIAO, K., CHEN, E., MA, H., AND HU, G. 2020. Learning or forgetting? a dynamic approach for tracking the knowledge proficiency of students. *ACM Transactions on Information Systems (TOIS) 38,* 2, 1–33.

KILICKAYA, F. ET AL. 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.

KINGMA, D. P. AND BA, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

LEE, D. AND SEUNG, H. S. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 535–541.

LI, J., WANG, F., LIU, Q., ZHU, M., HUANG, W., HUANG, Z., CHEN, E., SU, Y., AND WANG, S. 2022. Hiercdf: A bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 904–913.

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.

LIU, Q., WU, R., CHEN, E., XU, G., SU, Y., CHEN, Z., AND HU, G. 2018. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST) 9,* 4, 1–26.

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

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*, Y. Matsumoto and R. Prasad, Eds. The COLING 2016 Organizing Committee, 3245–3253.

MA, B., HETTIARACHCHI, G. P., AND ANDO, Y. 2022. Format-aware item response theory for predicting vocabulary proficiency. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, 695–700.

MA, B., HETTIARACHCHI, G. P., FUKUI, S., AND ANDO, Y. 2023a. Each encounter counts: Modeling language learning and forgetting. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 79–88.

MA, B., HETTIARACHCHI, G. P., FUKUI, S., AND ANDO, Y. 2023b. Exploring the effectiveness of vocabulary proficiency diagnosis using linguistic concept and skill modeling. In *Proceedings of the 16th International Conference on Educational Data Mining*, M. Feng, T. Käser, and P. Talukdar, Eds. International Educational Data Mining Society, 149–159.

MA, H., ZHU, J., YANG, S., LIU, Q., ZHANG, H., ZHANG, X., CAO, Y., AND ZHAO, X. 2022. A prerequisite attention model for knowledge proficiency diagnosis of students. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, 4304–4308.

MNIH, A. AND SALAKHUTDINOV, R. R. 2007. Probabilistic matrix factorization. In *Advances in neural information processing systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Vol. 20. Curran Associates, Inc.

NATION, I. S. 2001. *Learning vocabulary in another language*. Vol. 10. Cambridge university press Cambridge.

RECKASE, M. D. 2009. Multidimensional item response theory models. In *Multidimensional item response theory*. Springer, 79–112.

ROBERTSON, F. 2021. Word discriminations for vocabulary inventory prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, R. Mitkov and G. Angelova, Eds. INCOMA Ltd., 1188–1195.

SETTLES, B., T LAFLAIR, G., AND HAGIWARA, M. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics 8*, 247–263.

SONG, L., HE, M., SHANG, X., YANG, C., LIU, J., YU, M., AND LU, Y. 2023. A deep cross-modal neural cognitive diagnosis framework for modeling student performance. *Expert Systems with Applications*, 120675.

STÆHR, L. S. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal 36,* 2, 139–152.

SUN, Y., YE, S., INOUE, S., AND SUN, Y. 2014. Alternating recursive method for q-matrix learning. In *Proceedings of the 7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren, Eds. International Educational Data Mining Society, 14–20.

SUSANTI, Y., NISHIKAWA, H., TOKUNAGA, T., OBARI, H., ET AL. 2016. Item difficulty analysis of english vocabulary questions. In *Proceedings of the 8th International Conference on Computer Supported Education (CSEDU 2016)*, J. Uhomoibhi, G. Costagliola, S. Zvacek, and B. M. McLaren, Eds. Vol. 1. SCITEPRESS - Science and Technology Publications, Lda, 267–274.

THAI-NGHE, N. AND SCHMIDT-THIEME, L. 2015. Multi-relational factorization models for student modeling in intelligent tutoring systems. In *2015 Seventh international conference on knowledge and systems engineering (KSE)*. IEEE, 61–66.

TONG, S., LIU, J., HONG, Y., HUANG, Z., WU, L., LIU, Q., HUANG, W., CHEN, E., AND ZHANG, D. 2022. Incremental cognitive diagnosis for intelligent education. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 1760–1770.

TONG, S., LIU, Q., YU, R., HUANG, W., HUANG, Z., PARDOS, Z. A., AND JIANG, W. 2021. Item response ranking for cognitive diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence, 1750–1756.

TOSCHER, A. AND JAHRER, M. 2010. Collaborative filtering applied to educational data mining. *KDD Cup 2010 Workshop, Held as part of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*.

TRAUB, R. E. 1993. On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, W. C. Ward and R. E. Bennett, Eds. Routledge, 29–44.

VAN DER LINDEN, W. J. AND HAMBLETON, R. 1997. *Handbook of item response theory*. Vol. 1. Taylor & Francis Group.

VAN DER MAATEN, L. AND HINTON, G. 2008. Visualizing data using t-sne. *Journal of machine learning research 9,* 11.

WANG, F., LIU, Q., CHEN, E., HUANG, Z., CHEN, Y., YIN, Y., HUANG, Z., AND WANG, S. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. AAAI Press, 6153–6161.

WANG, F., LIU, Q., CHEN, E., HUANG, Z., YIN, Y., WANG, S., AND SU, Y. 2023. Neuralcd: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering 35,* 8, 8312–8327.

WANG, X., HUANG, C., CAI, J., AND CHEN, L. 2021. Using knowledge concept aggregation towards accurate cognitive diagnosis. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, 2010–2019.

WANG, Z., GU, Y., LAN, A., AND BARANIUK, R. 2020. Varfa: A variational factor analysis framework for efficient bayesian learning analytics. In *Proceedings of the 13th International Conference on Educational Data Mining*, A. N. Rafferty, J. Whitehill, V. Cavalli-Sforza, and C. Romero, Eds. International Educational Data Mining Society, 696–699.

YAO, L. AND 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.

ZYLICH, B. AND LAN, A. 2021. Linguistic skill modeling for second language acquisition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, 141–150.