Artificial intelligence-generated and human expert-designed vocabulary tests: a comparative study

Luo, Y., Wei, W., & Zheng, Y.

Artificial intelligence (AI) technologies have the potential to reduce the workload for the second language (L2) teachers and test developers. We propose two AI distractor-generating methods for creating Chinese vocabulary items: semantic similarity and visual similarity. Semantic similarity refers to antonyms and synonyms, while visual similarity refers to the phenomenon that two phrases share one or more characters in common. This study explores the construct validity of the two types of selected-response vocabulary tests (AI-generated items and human expert-designed items) and compares their item difficulty and item discrimination. Both quantitative and qualitative data were collected. 78 students from Beijing Language and Culture University were asked to respond to AI-generated and human expert-designed items respectively. Students' scores were analysed using the two-parameter item response theory (2PL-IRT) model. 13 students were then invited to report their test taking strategies in the think-aloud section. The findings from the students' item responses revealed that the human expert-designed items were easier but had more discriminating power than the AI-generated items. The results of think-aloud data indicated that the AI-generated items and expert-designed items might assess different constructs, in which the former elicited test takers' bottom-up test-taking strategies while the latter seemed more likely to trigger test takers' rote memorization ability.

Keywords: Vocabulary Test, Artificial Intelligence, Construct Validity, Computerised Test

Introduction

The gradually increasing population of Chinese language learners has created a growing demand for online Chinese vocabulary tests. Current debates on "assessment for learning" encourage test

designers, especially those in the field of computer assisted language learning (CALL) and mobile-assisted language learning (MALL), to create a user-friendly platform that better facilitates learners' self-assessment and self-reflection (Chen, Carger & Smith, 2017). Research in CALL and MALL demonstrates that Natural Language Processing (NLP) has the potential to assist language testing from at least two perspectives: automatic item generation (Chapelle & Chung, 2010) and automatic scoring (Voss, 2018). Automatic scoring techniques have already been applied in rating test takers' writing (Xi, Higgins, Zechner & Williamson, 2008), speaking (Voss, 2018), listening and reading abilities (Madsen, 1991), whereas the automatic item generation is frequently utilized in English vocabulary tests (Susanti, Tokunaga, Nishikawa, & Obari, 2017).

Although there are many studies on automatic test item generators for English language testing, the implications for testing Chinese as a second language are yet to be explored. Artificial Intelligence (AI) technologies, whichare receiving increased attention in the field of automatic vocabulary item generation, can fill this gap for three reasons (e.g., Susanti, Tokunaga, & Nishikawa, 2020; Ulum, 2020): firstly, both selected- and constructed- response formats can be generated with the application of NLP: (1) cloze items (Sakaguchi et al., 2013), (2) multiple-choice vocabulary items (e.g., Aldabe et al., 2006; Hoshino & Nakagawa, 2005), and (3) error correction items (e.g., Aldabe et al., 2006). Secondly, NLP technologies have the potential to create a larger number of distractors for multiple-choice questions (MCQs) in a short period of time using the four approaches as follows: (1) the corpus-based approach (e.g., Aldabe & Maritxalar, 2010), (2) the graph-based approach (e.g., Papasalouros et al., 2008), (3) Word2vec (e.g., Mikolov et al., 2013) and (4) visual similarity (e.g., Jiang & Lee, 2017). Among these four approaches, there are two promising language embedding tools to create semantically similar distractors: Word2vec and Latent Semantic Analysis

(LSA) in the corpus-based approach (a method to extract the meaning of a word based on the co-occurrence model) (Altszyler et al., 2017). Thirdly, different types of information resources can be used in NLP. Most of the empirical studies on vocabulary assessment have been conducted with existing corpora (Brown et al., 2005), but there is research utilizing other resources. For example, a real-time automatic MCQ generator designed by Hoshino and Nakagawa (2005) can extract important words and phrases from online articles to generate grammar and vocabulary items. Bearing in mind the possibilities of integrating NLP into the design of vocabulary test items, this study investigates the construct validity and compares the difficulty level and discrimination rate of two vocabulary tests which assess the same list of vocabulary items: AI-generated items and human expert-designed items in the context of learning Chinese as a foreign language.

Literature Review

The debates on the theoretical framework of the construct validity of vocabulary assessment have been continuing for years without a widely agreed definition. Schmitt et al. (2020) criticize current vocabulary tests for six reasons, among which three can be linked to this study: (1) an unspecified test purpose, (2) the generalization of intended test takers, and (3) the undefined aspects of assessed vocabulary knowledge. In this light, the definition of vocabulary ability should be specified before developing a test.

There are three distinct components defined by previous researchers on vocabulary ability: the first refers to the context of vocabulary use, which affects the lexical meaning in three ways: (1) differences across generations, (2) differences in interpretation across language varieties, and (3) differences in terminologies (Chapelle, 1994; Read, 2000). The second relates to vocabulary knowledge (Bruton, 2009; Chapelle, 1994). For example, Nation (2013) suggests there are various

levels of knowing a word: form (spoken form, written form, and word parts), meaning (form and meaning, concept and reference, and association) and use (grammatical functions, collocations, and constraints on use). All these three levels of vocabulary knowledge can be further classified as either receptive or productive vocabulary knowledge. Receptive vocabulary knowledge means that learners can recognize or comprehend the words, whereas productive vocabulary knowledge means that learners can use them in written or oral communications. The third component concerns the strategic competence in the use of words (Bachman, 1990; Chapelle, 1994). For example, according to taxonomies proposed by Gyllstad et al. (2015), there are six test-taking strategies in the context of selected-response items (e.g., MCQs): (1) knowing the meaning, (2) inferring the meaning from a known member of the word family (e.g., participants choose the correct answer *develop* because they know the meaning of *development*), (3) elimination, (4) inferring the meaning from similar words in test items, (5) inferring the meaning based on the context of the sentence and (6) blind guessing.

In aligning with the taxonomies, two processing models of word recognition are proposed: bottom-up and top-down (Færch & Kasper, 1987). Bottom-up processing starts from phonemes and morphemes to clauses and texts before linking to semantic content, whereas top-down processing works in the opposite direction (Matthew, 2014). Numbers of empirical studies related to vocabulary learning have indicated that bottom-up processing strategies are more effective than top-down strategies in facilitating English vocabulary learning (e.g., Barabadi & Khajavi, 2017; Makany et al., 2009; Moskovsky et al., 2015). Here bottom-up refers to learning vocabulary by associating words and smaller lexical units (e.g., morphemes), whereas top-down refers to the conventional way to learn vocabulary based on glossaries and outlines (Cairns et al., 1981). For instance, by inviting 120 students divided into two instruction groups, Moskovsky et al. (2015) concluded that the bottom-up

group outperformed the top-down one on both receptive vocabulary size and controlled productive vocabulary tests.

Five factors are reported as significant predictors of word processing models in vocabulary tests: (1) familiarity with the context (Ertürk & Mumford, 2017), (2) language proficiency (Haastrup, 2008), (3) test takers' experience of previous tests (Cohen, 2006), (4) item difficulty (Morimoto, 2007) and (5) the characteristics of the item response format (Cohen & Upton, 2006). For example, Ertürk and Mumford (2017) defined the familiarity of the context as the frequency of exposure to specific materials. By conducting a focus group interview, they found that learners seemed more likely to employ top-down strategies to process the texts they were familiar with. On the other hand, Haastrup (2008) investigated the relationship between learners' language proficiency and the use of test-taking strategies. The think-aloud data demonstrated that the high- and intermediate- level learners appeared to use top-down processing strategies more frequently than low-level learners. However, intermediate learners might change processing models based on the perceived difficulty of the context. Furthermore, Scouller and Prosser (1994) looked at the association between test performance and strategies, and found that test takers might reinforce a strategy which had helped them to succeed in previous tests. By inviting 21 participants to report their test-taking strategies in the think-aloud section, Morimoto (2007) concluded that learners might prefer top-down processing in easy multiple-choice vocabulary items. Finally, some researchers (e.g., Cohen & Upton, 2006) have reported that the characteristics of the item response format may be another predictor. For example, the mental processes in completing MCQ items are choice-oriented, which never occurs in the blank-filling items.

Apart from the two word-processing models, the selection of item types is another factor that needs careful consideration in the vocabulary test design. Multiple-choice questions have been used to assess the vocabulary knowledge at the level of form and meaning for a long time (Kremmel & Schmitt, 2016). However, most findings related to the item difficulty and discrimination of MCQs have been conducted in traditional paper-and-pen exam conditions. According to Cohen (2006) and Read (2000), there were at least four influential factors affecting item difficulty: (1) the item stem, (2) the distractors, (3) the targeted word, and (4) test-taking strategies. For example, Mori's study (2002) examined the success rate of inferring the word's meaning in three different situations: decontextualized items, contextualized items without options, and contextualized items with options. Her results had two implications: one was that contextualized items might differ in inferability from the context, and the other was that learners seemed likely to resort to options as clues. On the other hand, by assigning learners contextualized vocabulary tests with different distractors, Goodrich (1977) found that those distractors that were semantically related to the correct answer were more difficult than those that were similar to the correct answer in spelling (e.g., *beard* for *bread*). Moreover, in the context of Chinese vocabulary tests, McQueen (1996) concluded that three variables associated with the targeted word might affect the item difficulty: (a) the difficulty of pronunciation; (b) the frequency of exposure in learners' textbooks; (c) the form of presentation (i.e. the spoken or written form of Chinese). In addition to the three factors above, through a retrospective interview, Gyllstad et al. (2015) reported at least three test-taking strategies affecting item difficulty in multiple-choice vocabulary items: (a) elimination, (b) inferring the meaning from similar words in test items and (c) inferring the meaning based on the context of the sentence, since test takers who used these strategies might demonstrate no or partial vocabulary knowledge but they still arrived at

the correct answer. In contrast to research on item difficulty, to the best of our knowledge, studies on item discrimination have been scarce. Deane et al. (2014) have pointed out that item discrimination may be affected by a wide range of factors, such as distractors employed, the item type (e.g., topical associate), word frequency, etc.

In the field of MCQs' distractor generation, the applications of NLP include the following four approaches: (1) the corpus-based approach, which refers to the technique of extracting the meaning of the targeted word from those words that often co-occur with it in a corpus (Aldabe & Maritxalar, 2010); (2) the graph-based approach, which refers to the system using knowledge resources to calculate the semantic distance between two concepts or words (Liu et al., 2018); (3) Word2vec, which is a neural-network language embedding tool to predict the most semantically similar words of the targeted word (Mikolov et al., 2013); (4) the visual similarity method, which refers to the algorithm to search for distractors sharing at least one character in common with the targeted word in Chinese (Jiang & Lee, 2017). Briefly, the first three approaches aim at generating distractors that are semantically similar to the targeted word, whereas the last approach is to search for distractors that are visually similar to the targeted word. According to the first approach, two words that often co-occur with each other share high semantic similarity. For example, the word "red" and the word "ball" are semantically similar because they are near each other in the sentence "John holds a red ball." Aldabe and Maritxalar (2010) investigated the functionality of generated distractors with the LSA model, in which the semantic similarity between two words could be computed after vectorization. The distractors were thus generated due to the high semantic similarity in LSA, and the LSA distractors achieved a desirable result with 59% of them selected in the test involving 266 participants.

Secondly, as Papasalouros et al. (2008) point out, ontology is a fundamental resource in the graph-based approach, which refers to domain knowledge consisting of definitions of basic concepts, individuals, and relations between concepts and individuals. Their approach has three different strategies to generate MCQs' distractors: (1) class-based strategies following a hierarchy model (e.g., thing, technology, and ancient Greek technology), (2) property-based strategies following roles of individuals (e.g., *Polykrates* and *Eupalinos* belong to the class of **Person**, whereas *Herodotus* belongs to its subclass of **Historian.** Thus, "*Polykrates* **hired** *Eupalinos*" is the correct answer but "*Polykrates* **hired** *Herodotus*" is a distractor) and (3) terminology-based strategies following concept/sub-concept relationships (e.g., "*Herodotus* is a **historian**" is a correct answer since **Herodotus** is a subclass of **Historian.** "**Politician** is a **historian**" is a distractor since **Politician** is a sibling class of **Historian**"). They invited two educational experts to evaluate the generated items about the Greek Eupalineio Tunnel ontology from three perspectives: pedagogical quality, syntactic correctness and the number of generated questions. The generated questions were generally satisfactory but not all items were syntactically correct, since 22 out of 88 generated items did not make any sense.

Thirdly, regarding the Word2vec method, Mikolov et al. (2013) mention two neural network language models in Word2vec: Continuous Bag of Words (CBOW) and Skip-gram. In the CBOW model, surrounding words are combined to predict the output word in the middle; while on the other hand, in Skip-gram, the output words can be predicted by one input word. The Skip-gram model is widely used to capture semantically similar words as it can predict each targeted word's context. For example, for the targeted word "ball," the Skip-gram model is likely to associate it with the word "bouncy" rather than the word "red" because the word "red" may co-occur with many other words

and “bouncy” is more discriminative of the targeted word’s context (Hollis et al., 2017). According to Altszyler et al. (2017), a word can be converted to a vectorial value in a toolkit named Gensim, and the semantic similarity (S) between two random words can be measured as follows:

$$S(v_1, v_2) = \cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\| v_1 \| \cdot \| v_2 \|}$$

Finally, regarding the visual similarity method, Jiang and Lee (2017) note that this generation method can be used in designing Chinese vocabulary assessment. For example, “爱好” (hobby) and “爱情” (love), two Chinese phrases that have different meanings but share the same first character “爱”, can be plausible for learners.

To elaborate the strengths of the four approaches, there are three main advantages as follows: firstly, corpus-based and Word2vec approaches can quickly generate semantically related distractors. In particular, the Word2vec method has the advantage of processing a large corpus, whereas the corpus-based approach performs better in a small size corpus (Altszyle et al., 2017). Secondly, the graph-based approach can assess domain knowledge and terminologies (Papasalouros et al., 2008). Thirdly, the visual similarity approach has the potential of generating numbers of orthographical distractors in a short time with a corpus analysis tool (Jiang & Lee, 2017).

On the other hand, to elaborate the limitations of NLP techniques, the four approaches have their own disadvantages. The corpus-based approach can only assess superficial knowledge rather than professional knowledge (Alsubait et al., 2013). By contrast, the graph-based approach can only assess specific domain knowledge and it needs a more sophisticated algorithm to improve its performance on the syntactic correctness of generated items (Papasalouros et al., 2008). Likewise, the Word2vec is likely to generate semantically incorrect distractors with a small size corpus (Altszyle et al., 2017). The visual similarity method has two limitations: one is that the generated

distractors may not be as attractive as those distractors that are semantically related to the correct answer, and the other is that this method is restricted to searching for Chinese characters (Jiang & Lee, 2017).

We used the Skip-gram model in Word2vec and the visual similarity method as the generation algorithms for MCQs' distractors. This study contributes to the knowledge and practice of vocabulary assessment in the context of second language learning from at least two perspectives:

(1) Our study investigates human-designed and AI-designed vocabulary assessment tasks from the perspectives of item difficulty, item discrimination and construct validity.
(2) We validate an NLP approach to generating vocabulary tests, which has not been sufficiently investigated in previous vocabulary assessment studies.

Research Design

Since this study investigates the validity of newly designed AI-generated Chinese vocabulary test items and human expert-designed test items, it asks two research questions:

RQ1: With other factors controlled (e.g., language proficiency of test takers), are AI-generated items more difficult and have a higher discrimination rate than those designed by human experts?

RQ2: Do AI-generated items lead test takers to use different test-taking strategies in comparison with items designed by human experts?

In this study, there are four stages of test construction and validation (see Figure 1): (1) the construction of AI-generated items, (2) the construction of expert-designed items, (3) the comparison between the two item types in terms of two item features (item difficulty and item discrimination in the 2PL-IRT model), and (4) the construct validation of the two item types with the think-aloud

protocol. There are four main considerations when designing an AI-generated vocabulary test: (1) the selection of the corpus, (2) the selection of keywords/phrases for testing, (3) the response formats, and (4) the methods of generating distractors. Firstly, the Sogou News corpus (https://www.sogou.com/labs/resource/cs.php) was chosen due to the large amount of textual information (420,000 items of news). The two other free corpora, Wikipedia Chinese version and the Chinese subtitles corpus of movies and drama (https://github.com/warmheartli/ChatBotCourse/tree/master/), have their limitations: the Wikipedia Chinese version has too many difficult sentences with complicated grammatical structures, and the Chinese subtitles corpus of movies and dramas contains too many oral expressions and fragmented syntactic structures. Afterwards, a word processing software named jieba (Sun, 2012) was used to segment Chinese sentences to create space between words like English and Spanish, which was the pre-processing procedure for the further calculation of semantic similarity between words.

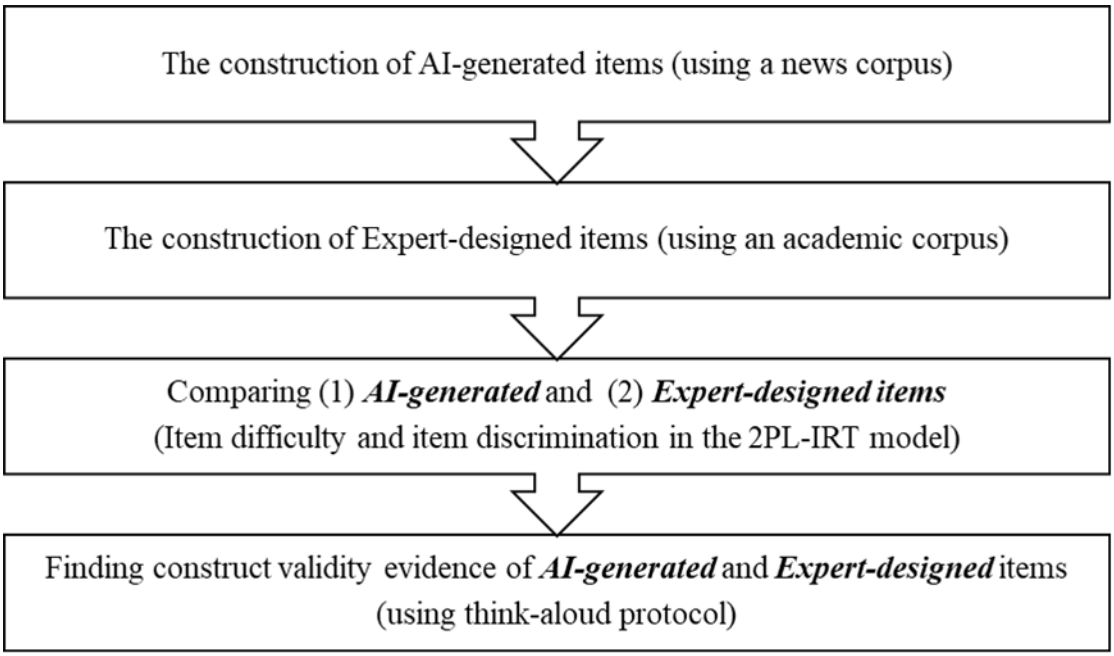

Figure 1. The flowchart of this study

Secondly, 30 content words (6.3% of the HSK level 4 content words) were selected as the keywords to generate items for three reasons (see Appendix A): first, the HSK exam system has been the most popular Chinese proficiency test. Second, in our study, students from HSK level 4 were the most representative cohort of students among three groups of participants (29 from HSK level 3, 24 from HSK level 4, 25 from HSK level 5). Third, according to Singleton's (1999) suggestion, we focused on content words (e.g., nouns, verbs, and adjectives), and selected 30 content words (mostly compound words) at HSK level 4 as keywords for distractor generation.

Thirdly, four response formats were generated with the AI-generator: two item formats to select the most appropriate answer with or without context (CA: contextualized items asking test takers to select the most appropriate option; DA: decontextualized items asking test takers to select the most appropriate option), and the other two to select the most inappropriate answer with or without context (CI: contextualized items asking test takers to select the most inappropriate option; DI: decontextualized items asking test takers to select the most inappropriate option). The keys of these four item types were different: the keys of CA and DA were the targeted words in the HSK vocabulary curriculum, whereas the keys of CI and DI were the alternatives of the distractors extracted from CA and DA (e.g., the antonym of the targeted word). Meanwhile, two factors were taken into consideration before writing stems: first, an item stem should be limited in the range of 10-20 words to provide sufficient contextual clues for test takers. Second, an item stem should consist of one or two simple declarative sentences. Therefore, after automatic selection and manual adjustment, the contextualizing sentences had no more than 20 characters to ensure semantic simplicity and integrity.

Fourthly, for distractor generation, two generating methods were selected: semantic similarity (Mikolov et al., 2013) and visual similarity of the Chinese characters (Jiang & Lee, 2017). Semantic similarity refers to antonyms and synonyms, while visual similarity refers to the phenomenon that two phrases share one or more characters in common. According to the semantic and visual closeness with the targeted word, all generated distractors were divided into a high level group and a low level group (see Table 1). Taking "爱好" as an example, the distractors can be "爱心" (Low semantic similarity + High visual similarity), "天分" (Low semantic similarity + Low visual similarity) and "厌恶" (High semantic similarity + Low visual similarity) (see Appendix B). It is worth noting that the antonyms of the targeted word cannot be automatically generated. Thus, we used a reputable Chinese dictionary, Online Xinhua Dictionary, as a reference to find corresponding antonyms.

Table 1. The outline of AI-generated items

Group	Explanation	Distractors
CA	Context + the most appropriate option	1. Low semantic + High visual similarity 2. Low semantic + Low visual similarity 3. High semantic + Low visual similarity
DA	No context + the most appropriate option	1. Low semantic + High visual similarity 2. Low semantic + Low visual similarity 3. High semantic + Low visual similarity
CI	Context + the most inappropriate option	1. Semantic High 1 2. Semantic High 2 3. Semantic High 3
DI	No context + the most inappropriate option	1. Semantic High 1 2. Semantic High 2

	3. Semantic High 3

Data collection

To answer the first research question, two experienced teachers from BLCU (Beijing Language and Culture University) were invited to create 30 MCQs based on their teaching experiences and materials (see Appendix C). Beijing Language and Culture University is one of the universities with a large number of foreign students learning Chinese as a foreign language. The expert-designed items met two criteria: one was that those items designed by experts were contextualized items extracted from the teaching materials, which were similar to the CA items in the AI-generated items to a certain extent. The other was that the alternatives of distractors were mainly high-frequency words in the textbooks, which test takers were very familiar with. In general, the 4 AI-generated item types and 1 expert-designed item type shared one common core, that is, to assess knowledge of the targeted words at HSK level 4 (e.g., form and meaning, and constraint). In total, 78 participants from BLCU were invited to complete 150 MCQs with five question types. These 78 participants included students at different levels, and the minimum number of every subgroup was more than 20 (29 at HSK level 3, 24 at HSK level 4 and 25 at HSK level 5). Most students were from Bangladesh, Thailand, South Korea, Russia, and Pakistan (see Table 2). They were first asked to respond to the MCQs via their mobile phones within a time limit of 90 minutes. To answer the second research question, some participants were further invited to report their test-taking strategies using think-aloud protocols in the second section after two days. Using the think-aloud protocol as part of language assessment validation evidence is a common practice in this field (Gyllstad et al., 2015; Nation, 2013). Data collection in the think-aloud section consisted of two phases. In Phase 1, all 13

invited students from three different levels attended a 30-minute training session on how to think aloud when answering items. In Phase 2, they were asked to share the processing strategies they used to arrive at their answers, in which 25 items (5 randomly selected items from each item type) were selected to trigger their processing strategies. Each participant was questioned as soon as one item was finished, and the following probes were used to minimize the long silence: "What are you thinking of? Why do you pick this answer?" The think-aloud section was audio-recorded and conducted in Chinese.

Table 2. **Demographic information of test takers**

The MCQs section's participants	Demographics	Percentage	Number
Gender:	Female	62.8%	49
	Male	37.2%	29
Proficiency:	Level 3	37.2%	29
	Level 4	30.8%	24
	Level 5	32.0%	25
Native language:	Bangladesh	16.7%	13
	Thailand	10.3%	8
	South Korea	10.3%	8
	Others	62.7%	49
The think-aloud section's participants			
Gender:	Female	53.8%	7
	Male	46.2%	6
Proficiency:	Level 3	38.5%	5
	Level 4	38.5%	5
	Level 5	23.0%	3

Data analysis

To answer the first research question, we used the two-parameter logistic (2PL) model in item response theory (IRT) to calculate item difficulty and item discrimination. IRT has been recognized as one of the most appropriate psychometric models in accounting for item features (difficulty, discrimination, guessing). Apart from its advantage of handling both large and small sample sizes, it can also be specialized for handling missing data (Cai & Kunnan, 2018). Before estimating the two parameters, we first examined each item's outfit significant p-value. The item outfit significant p-value is an indicator of mis-functioning items that should be eliminated. When an item violates its lower limitation ($p = 0.05$), it means this item is either too easy or difficult for learners. The 2PL model uses item difficulty (ranging from -3 to 3) and item discrimination (ranging from 0 to 3) to estimate the fit between item quality and person ability (Adams et al., 1997; Boone, 2016; DeMars, 2010). Compared with low-proficiency learners, high-proficiency learners are more likely to arrive at the correct answer with increased levels of the two parameters. All parameters were estimated using the R software and the TAM package. (Robitzsch et al., 2020).

To answer the second research question, as Table 3 suggests, there are four categories of test-taking strategies that may be used by learners in previous studies (Gyllstad et al., 2015; Nation, 2013). In addition, considering that there is no word part in Chinese, the code associated with "word parts" has been changed to the "strategy based on radicals of a character", because a Chinese character normally has two radicals and one of them contains the lexical meaning (Liu et al., 2018).

Table 3. The coding scheme's outline and examples

Strategies	Examples:
Memorization Strategy	

Based on learned example sentences	Our teacher always says “don’t give up” and “try as hard as possible.” Therefore, I choose this option.
Lexical Information Strategies	
Based on a word’s meaning	Only the option D “cease” (终止) means to stop. So, I choose this one.
Based on a word’s meaning in the stem	Because of the word “save”, option C “damage” (破坏) cannot be selected. “Save” is a positive word, while “damage” is negative.
Based on a morpheme in a word	I choose the option B “advantage” (优点), because of the character “点”. This word seems to be related to math or so.
Based on a word collocation	Option D “lack” (缺乏) is different from the other three options, since there should be a stuff followed.
Based on part of speech	I think that three options are not correct, as there should be an adjective connected with the aspect marker “di” (地).
Based on radicals of a character	The right part of the character (掉) is very similar with the character “table” (桌).
Context Based Strategy	
Based on the sentence’s meaning	This sentence could happen between two people. Thus, there should be another sentence like “what kind of house would you like” in the front of this one.
Blind Guessing	
Blind Guessing	I don’t know the meaning of option A. So, I draw it for good.

Findings

At the beginning of the analysis, all items were estimated for the outfit significance p-value. The results of preliminary estimation revealed that 9 items violated the outfit p-value’s lower limit (0.05), in which 8 words were involved: CI item 1, DA item 1, CI item 2, Expert item 3, CA item 15, Expert item 17, Expert item 21, Expert item 23, and DA item 29. These 8 words and their corresponding item types were removed. It is worthy of note that these items that violated the lower boundary of the

outfit p-value were also lower than the acceptable range of item difficulty. All these removed keywords had set phrases test takers were familiar with, which might be the interference factor of the outfit p-value and item difficulty. Eventually, 22 out of 30 keywords and corresponding item types were kept for further analyses. Table 4 shows the descriptive statistics of the two item features in the AI-generated and expert-designed items.

Table 4. Descriptive data of five item types in item difficulty and item discrimination

	Explanations	Mean	SD	Skew
Difficulty				
CI	Contextualized MCQ + select the most inappropriate option	-0.52	0.72	-0.21
CA	Contextualized MCQ + select the most appropriate option	-0.93	1.05	0.01
DA	Decontextualized MCQ + select the most appropriate option	-1.05	0.92	0.81
DI	Decontextualized MCQ + select the most inappropriate option	-1.07	1.15	0.19
Expert	Contextualized MCQ + select the most appropriate option	-1.28	1.07	0.35
Discrimination				
Expert	Contextualized MCQ + select the most appropriate option	1.67	0.69	0.63
CA	Contextualized MCQ + select the most appropriate option	1.52	0.72	-0.02
DI	Decontextualized MCQ + select the most inappropriate option	1.48	0.81	0.33
DA	Decontextualized MCQ + select the most appropriate option	1.36	0.57	0.75
CI	Contextualized MCQ + select the most inappropriate option	1.09	0.59	0.33

As for the first research question, firstly, the AI-generated items were descriptively more difficult than those items designed by experts with no significant difference according to the results of ANOVA, $F(4,105) = 1.76$, $p = 0.14$. The CI and CA were the lowest among the five item types in terms of item difficulty, suggesting that test takers might have the worst performance when they

utilized contextual clues extracted from the news corpus. Table 5 demonstrates the ANOVA results of five item types in terms of item difficulty.

Table 5. Comparison of five item types in item difficulty

Group	N	Mean	SD	F	*p*
CA (a)	22	-0.93	1.05	1.76	0.14
CI (b)	22	-0.52	0.72		
DA(c)	22	-1.05	0.92		
DI (d)	22	-1.07	1.15		
Expert (e)	22	-1.28	1.07		

Then, regarding item discrimination, expert-designed items had more discriminating power than AI-generated items, and there was no significant difference between them according to the result of ANOVA, $F(4,105) = 2.19$, $p = 0.075$. To be more specific, the discriminating power ranked from the best to least as follows: Expert, CA, DI, DA, and CI. Table 6 demonstrates the ANOVA results of five item types in terms of item discrimination.

Table 6. Comparison of five item types in item discrimination

Group	N	Mean	SD	F	*p*
CA (a)	22	1.52	0.72	2.19	0.08
CI (b)	22	1.09	0.59		
DA(c)	22	1.36	0.57		
DI (d)	22	1.48	0.81		
Expert (e)	22	1.67	0.69		

The second research question addressed two aspects of the think-aloud data: the diversity and the pattern of using test-taking strategies. Table 7 shows the differences between AI-generated and

expert-designed items in terms of test-taking strategies' diversity. As to the diversity, the differences between AI-generated items and expert-designed items existed in four categories: firstly, the memorization strategy was more frequently used in the expert-designed items than AI-generated items; secondly, the participants paid more attention to radicals, morphemes (characters) and words in AI-generated items compared with expert-designed items; thirdly, the context-based strategy was more frequently used in expert-designed items than AI-generated items; fourthly, the blind guessing strategy was more frequently used in AI-generated items than expert-designed items, especially for CA. In addition, it is worth noting that the participants at HSK level 4 tended to use more lexical information strategies in AI-generated items, whereas they used the memorization and context-based strategies more frequently in expert-designed items.

Moreover, it seemed that participants processed the language messages with different processing models: with items produced by the experts, participants tended to use the top-down processing model, since those smaller components, such as radicals, morphemes and words, received less attention than contextual clues. However, with AI-generated items, the participants paid more attention to the word and other smaller components, which was close to the bottom-up processing model.

The different resources of test-writing might be the underlying factor affecting the diversity of test-taking strategies. In the condition of AI-generated items, the item stem was extracted from a news corpus these test takers were not familiar with. Therefore, bottom-up processing strategies were applied to perceive the meaning of the lexical unit before linking it to the concept in the AI-generated item. On the other hand, with items produced by human experts, participants used the top-

down processing model to recall their learned sentences since sentences extracted from the teaching materials might trigger test takers' memory.

Table 7. Test-taking strategies' diversity of five item types

Strategies	**Item types**				
	CA	CI	DA	DI	Expert
Memorization Strategy					
Based on learned example sentences in memory	8	7	4	4	10
Lexical Information Strategies					
Based on a certain word's meaning	28	25	41	40	19
Based on a certain word's meaning in the stem	15	4	0	0	5
Based on a morpheme in a compound word	1	8	5	7	3
Based on a word collocation	1	1	0	0	0
Based on part of speech	1	1	0	0	1
Based on radicals of a certain character	4	6	5	3	2
Context Based Strategy					
Based on the sentence's meaning	16	3	0	0	17
Blind Guessing					
Guessing based on no ground	8	1	2	3	1

Secondly, Table 8 shows the differences between AI-generated and expert-designed items in terms of test-taking strategy patterns. In total, two differences were identified: first, compared with expert-designed items, participants tended to start from analysing the meaning of a certain word in AI-generated items. Second, participants were likely to perceive the expert-designed items from the memorization strategy, whereas this tendency was not observed in AI-generated item types. These findings were also consistent with the results of diversity, that is, the test-taking pattern in AI-

generated items appeared to be bottom-up and the pattern in expert-designed items seemed to be the opposite. These findings suggested that the different resources of item-writing might account not only for the diversity of test-taking strategies, but also the patterns of test-taking strategies.

Table 8. Test-taking strategy patterns of five item types

	The most frequently reported pattern	The second most frequently reported pattern
CA	1. Understand the sentence's meaning 2. Arrive the answer (5 times)	1. Guess the sentence's meaning based the stem's keyword 2. Remove the option based on its meaning 3. Know the meaning of the option 4. Arrive at the answer (3 times)
CI	1. Remove the option based on its meaning 2. Know the meaning of the option 3. Arrive at the answer (4 times)	1. Know the meaning of the option 2. Arrive at the answer (4 times)
DA	1. Know the targeted word 2. Arrive at the answer (7 times)	1. Remove the option based on its meaning 2. Remove the option based on its meaning 3. Remove the option based on its meaning 4. Arrive at the answer (7 times)
DI	1. Know the targeted word 2. Remove the option based on its meaning 3. Arrive at the answer (11 times)	1. Know the meaning of the option 2. Arrive at the answer (8 times)
Expert	1. Understand the sentence's meaning 2. Arrive at the answer (13 times)	1. Recall learned sentences in memory 2. Arrive at the answer (9 times)

Discussion

This study investigated item difficulty, item discrimination and construct validity of both AI-generated items and expert-designed items, by collecting MCQ answers with five formats of vocabulary assessment (4 AI-generated item types and 1 expert-designed item type) from 78 participants and think-aloud data from 13 participants. By applying the 2PL-IRT model, it was found that AI-generated items are relatively more difficult than expert-designed items, but less able to discriminate the high- and low- performing test takers in comparison with expert-designed items. As for construct validity, the AI-generated and expert-designed items seem to assess different constructs due to their differences in test-taking strategies' diversity and patterns. Thus, the AI-generated items seem to assess the receptive vocabulary ability (e.g., using the context/sentence to make sense of vocabulary), whereas the expert-designed items assess the extent to which test takers can memorize the teachers' examples which include the targeted words.

Firstly, the IRT results demonstrate that the AI-generated items are descriptively more difficult than expert-designed items, which may be explained by the three predictors indicated in the literature part: (1) inferability from the context, (2) the difficulty of distractors and (3) the difficulty of the targeted word. Firstly, the think-aloud data demonstrate that contextual clues are utilized in both CA and expert-designed items with a similar frequency, whereas participants seem to use more lexical information strategies in CA compared with expert-designed items. In other words, unlike the expert-designed items, participants seem unlikely to infer the targeted word from the contextual clues provided by AI-generated items, which echoes Mori's (2002) conclusion that the degree of inferability from the context may affect the difficulty level. Secondly, compared with expert-designed items, AI-generated items appear to have more plausible distractors. For example, the AI item-generator may create distractors which are difficult to eliminate based on the first impression,

since those distractors are semantically related to each other (Gyllstad et al., 2015). This offers further support for Goodrich's (1977) conclusion that the difficulty of distractors may be an underlying factor affecting the difficulty level. Thirdly, there are some AI-generated items whose targeted words may not be presented in the textbook (e.g., the antonym of the targeted word), which supports the conclusion that targeted words which have not been presented in the textbook may be more difficult for test takers (McQueen, 1996).

Secondly, the results of item discrimination indicate that the expert-designed items have more discriminating power than AI-generated items. As Deane and other researchers (2014) mentioned, vocabulary "item types that required greater depth of semantic knowledge would tend to show greater difficulty and discrimination" (p. 1). This study further defines the depth of semantic knowledge in receptive vocabulary items from three perspectives: (1) contextual clues (whether the item provides contextual clues or not), (2) the similarity between the correct answer and distractors (semantically similar or visually similar), and (3) the item type (to select the appropriate or inappropriate answer).

Thirdly, the findings of think-aloud data indicate that the AI-generated items and expert-designed items may assess different constructs based on the evidence of test-taking strategies' patterns and diversity. The reasons why AI-generated and expert-designed items differ widely in test-taking patterns can be attributed to the following three factors: (1) familiarity with the context (Ertürk & Mumford, 2017), (2) language proficiency (Haastrup, 2008) and (3) experiences of prior tests (Cohen, 2006). Firstly, as indicated in the think-aloud data, the two experts used instruction materials as the resource to develop the vocabulary test, which created familiarity for the test takers to employ top-down strategies. This result echoes Ertürk and Mumford's (2017) conclusion that test

takers may employ top-down processing strategies in the context they are familiar with. Secondly, as demonstrated by the think-aloud data, despite the success of applying top-down strategies in expert-designed items, the intermediate participants seldom rely on top-down processing strategies in AI-generated items. In other words, intermediate learners may switch to bottom-up strategies in difficult items, which supports Haastrup's (2008) conclusion that intermediate learners cannot use the top-down processing as adeptly as high-level learners. Thirdly, even though context familiarity and language proficiency may be the dominant factors of the test-taking patterns, the think-aloud data reveal that there are still some participants perceiving an item from a specific strategy which they believe as useful. For example, Participant 6, whose language proficiency was HSK level 4, reported: "what I focus is the word collocation rather than the meaning of the sentence, since this strategy is time-saving and it has helped me to achieve desirable results in many vocabulary tests." This finding of test-taking patterns appears to support Cohen's (2006) idea about the impact of experiences on the patterns, which may explain the individual difference in test-taking patterns.

Regarding the diversity of processing strategies between AI-generated and expert-designed items, we propose two main reasons: item difficulty (Morimoto, 2007) and characteristics of the item response format (Cohen & Upton, 2006). Firstly, according to the difficulty indices calculated by the 2PL-IRT model, AI-generated items are descriptively more difficult than expert-designed items. As demonstrated by the think-aloud data, the context-based strategy is more frequently used in the easier item type. By contrast, for the more difficult item type, the context-based strategy has received less attention than lexical information strategies. This appears to further support Morimoto's (2007) conclusion that learners appear to make full use of contextual clues in easy items. Secondly, think-aloud data offer support to the conclusion that the mental process in MCQs is choice-oriented

(Cohen & Upton, 2006), since participants may take lexical information strategies as a resort to eliminate distractors in the condition of lacking knowledge of distractors or obtaining partial knowledge of the targeted word. For example, as demonstrated by the think-aloud data, radicals and morphemes have been used as clues to eliminate plausible AI-generated distractors, which is not surprising considering that test takers have not been frequently exposed to the news materials.

Moreover, the difference in test-taking patterns also provides an implication for vocabulary learning. According to Moskovsky et al. (2015), the bottom-up vocabulary learning approach is more effective than the top-down learning approach because the bottom-up processing of words may be more salient in learners' memory. Presumably, the AI-generated items may be better in facilitating vocabulary learning than the expert-designed items since they are more likely to trigger learners' bottom-up processing of words.

Conclusion

AI-generated vocabulary items have the potential to be a valid and reliable instrument to test candidates' Chinese vocabulary knowledge. This study has examined four receptive AI-generated vocabulary item types that can be generated in a very short time. The rigorous evidence of construct validity of our test has emerged in the think-aloud data, demonstrating that our test serves its purpose to assess the construct of form and meaning of receptive vocabulary knowledge.

In spite of these advantages, this study has at least two limitations: firstly, despite the promising performance in the 2PL-IRT model, the sample size is relatively small. Secondly, both semantic similarity and visual similarity methods may produce nonsense words and the semantic accuracy of distractors should be improved. In future research, we hope to use the AI algorithm to create other forms of vocabulary tests (e.g., productive vocabulary tests) and investigate the relationship between

generated MCQs and productive vocabulary tests. Furthermore, in the 2PL-IRT model, the guessing parameter is not considered; it is argued that not only the discrimination parameter but also the guessing parameter have fundamental roles in multiple-choice vocabulary items (Tseng, 2013). In the near future, more studies and relevant work will be conducted using the 3PL-IRT model to estimate learners' guessing rate in multiple-choice vocabulary items.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Aldabe, I., de Lacalle, M. L., Maritxalar, M., Martinez, E., & Uria, L. (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 584–594). Springer Berlin Heidelberg.

Aldabe, I., & Maritxalar, M. (2010). Automatic Distractor Generation for Domain Specific Texts. In H. Loftsson, E. Rögnvaldsson, & S. Helgadóttir (Eds.), *Advances in Natural Language Processing* (pp. 27–38). Springer Berlin Heidelberg.

Alsubait, T., Parsia, B., & Sattler, U. (2013, September 23-25). *A similarity-based theory of controlling MCQ difficulty* [Paper presentation]. 2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE), Lodz, Poland. https://doi.org/10.1109/ICeLeTE.2013.6644389

Altszyler, E., Ribeiro, S., Sigman, M., & Slezak, D. F. (2017). The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text. *Consciousness and*

Cognition, *56*, 178–187. https://doi.org/10.1016/j.concog.2017.09.004

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Barabadi, E., & Khajavi, Y. (2017). The effect of data-driven approach to teaching vocabulary on Iranian students' learning of English vocabulary. *Cogent Education*, *4*(1), 1-13. https://doi.org/10.1080/2331186X.2017.1283876

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sciences Education*, *15*(4), 1-7. https://doi.org/10.1187/cbe.16-04-0148

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005, October 6-8). *Automatic Question Generation for Vocabulary Assessment* [Paper presentation]. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada. https://www.aclweb.org/anthology/H05-1103.pdf

Bruton, A. (2009). The Vocabulary Knowledge Scale: A Critical Analysis. *Language Assessment Quarterly*, *6*(4), 288–297. https://doi.org/10.1080/15434300902801909

Cai, Y., & Kunnan, A. J. (2018). Examining the inseparability of content knowledge from LSP reading ability: An approach combining bifactor-multidimensional item response theory and structural equation modeling. *Language Assessment Quarterly*, *15*(2), 109–129. https://doi.org/10.1080/15434303.2018.1451532

Cairns, H. S., Cowart, W., & Jablon, A. D. (1981). Effects of prior context upon the integration of lexical information during sentence processing. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 445–453. https://doi.org/10.1016/S0022-5371(81)90551-X

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157–187. https://doi.org/10.1177/026765839401000203

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301-315. https://doi.org/10.1177/0265532210364405

Chen, Y., Carger, C. L., & Smith, T. J. (2017). Mobile-assisted narrative writing practice for young English language learners from a funds of knowledge approach. *Language Learning & Technology*, *21*(1), 28-41.

Cohen, A. D. (2006). The Coming of Age of Research on Test-Taking Strategies. *Language Assessment Quarterly*, *3*(4), 307–331. https://doi.org/10.1080/15434300701333129

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph Series Report No. 33). ETS. https://doi.org/10.1002/j.2333-8504.2006.tb02012.x

Deane, P., Lawless, R., Li, C., Sabatini, J., Bejar, I., & O'Reilly, T. (2014). Creating Vocabulary Item Types That Measure Students' Depth of Semantic Knowledge. *ETS Research Report Series*, *2014*(1), 1–19. https://doi.org/10.1002/ets2.12001

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Ertürk, N., & Mumford, S. E. (2017). Understanding test-takers' perceptions of difficulty in EAP vocabulary tests: The role of experiential factors. *Language Testing*, *34*(3), 413–433. https://doi.org/10.1177/0265532216673399

Færch, C., & Kasper, G. (Eds.). (1987). *Introspection in second language research*. Multilingual Matters.

Goodrich, H. C. (1977). Distractor Efficiency in Foreign Language Testing. *TESOL Quarterly*, *11*(1), 69-78. https://doi.org/10.2307/3585593

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), 278–306. https://doi.org/10.1075/itl.166.2.04gyl

Haastrup, K. (2008). Lexical Inferencing Procedures in Two Languages. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and Writing in a First and Second Language* (pp. 67–111). Palgrave Macmillan UK.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, *70*(8), 1603-1619. https://doi.org/10.1080/17470218.2016.1195417

Hoshino, A., & Nakagawa, H. (2005, June 29). *A Real-Time Multiple-Choice Question Generation For Language Testing: A Preliminary Study* [Paper presentation]. Proceedings of the Second Workshop on Building Educational Applications Using NLP, Ann Arbor, Michigan, United States. https://www.aclweb.org/anthology/W05-0203.pdf

Jiang, S., & Lee, J. (2017, September 8). *Distractor Generation for Chinese Fill-in-the-blank Items* [Paper presentation]. Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark. https://doi.org/10.18653/v1/W17-5015

Kremmel, B., & 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. https://doi.org/10.1080/15434303.2016.1237516

Liu, M., Rus, V., & Liu, L. (2018). Automatic Chinese Multiple Choice Question Generation Using Mixed Similarity Strategy. *IEEE Transactions on Learning Technologies*, *11*(2), 193–202. https://doi.org/10.1109/TLT.2017.2679009

Madsen, H. S. (1991). Computer-Adaptive Testing of Listening and Reading Comprehension: The Brigham Young University Approach. In P. Dunkel (Ed.), *Computer Assisted Language Learning and Testing: Research Issues and Practice* (pp. 237-257). Newbury House.

Makany, T., Kemp, J., & Dror, I. E. (2009). Optimising the use of note-taking as an external cognitive aid for increasing learning. *British Journal of Educational Technology*, *40*(4), 619–635. https://doi.org/10.1111/j.1467-8535.2008.00906.x

Matthews, P. H. (2014). *The concise Oxford dictionary of linguistics*. Oxford University Press.

McQueen, J. (1996). Rasch scaling: How valid is it as the basis for content-referenced descriptors of test performance? *Australian Review of Applied Linguistics. Supplement Series*, *13*(1), 137-187. https://doi.org/10.1075/aralss.13.07mcq

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, May 2-4). *Efficient estimation of word representations in vector space* [Paper presentation]. Proceedings of the Workshop at the International Conference on Learning Representations, Scottsdale, Arizona, United States.

Mori, Y. (2002). Individual differences in the integration of information from context and word parts in interpreting unknown kanji words. *Applied Psycholinguistics*, *23*(3), 375–397. https://doi.org/10.1017/S0142716402003041

Morimoto, Y. (2007). Test-taking Processes of Vocabulary Tests in Context From the Perspective of Think-Aloud Analysis. *JLTA Journal Kiyo*, *10*, 68–87. https://doi.org/10.20622/jltaj.10.0_68

Moskovsky, C., Jiang, G., Libert, A., & Fagan, S. (2015). Bottom-Up or Top-Down: English as a Foreign Language Vocabulary Instruction for Chinese University Students. *TESOL Quarterly*, *49*(2), 256–277. https://doi.org/10.1002/tesq.170

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Papasalouros A., Kotis K., & Kanaris K. (2008, July 22-25). *Automatic generation of multiple-choice questions from domain ontologies* [Paper presentation]. IADIS International Conference E-Learning 2008, Amsterdam, Netherlands.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Robitzsch, A., Kiefer, T., & Wu, M. (2020, March 3). *TAM: Test analysis modules. R package version 3.4-26*. https://CRAN.R-project.org/package=TAM

Sakaguchi, K., Arase, Y., & Komachi, M. (2013, August 4-9). *Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners* [Paper presentation]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria. https://www.aclweb.org/anthology/P13-2043.pdf

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

Scouller, K. M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, *19*(3), 267–279. https://doi.org/10.1080/03075079412331381870

Singleton, D. M. (1999). *Exploring the second language mental lexicon*. Cambridge University Press.

Sun, J. (2012, March 3). *Jieba Chinese word segmentation tool*. https://github.com/fxsjy/jieba/

Susanti, Y., Tokunaga, T., & Nishikawa, H. (2020). Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning*, *15*(1), 9. https://doi.org/10.1186/s41039-020-00132-w

Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2017). Controlling item difficulty for automatic

vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, *12*(1), 25. https://doi.org/10.1186/s41039-017-0065-5

Tseng, W. T. (2013). Validating a Pictorial Vocabulary Size Test via the 3PL-IRT Model. *Vocabulary Learning and Instruction*, *2*(1), 64-73.

Ulum, Ö. G. (2020). A critical deconstruction of computer-based test application in Turkish State University. *Education and Information Technologies*, *25*(6), 4883–4896. https://doi.org/10.1007/s10639-020-10199-z

Voss, E. (2018). Technology and Assessment. In J. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–7). Wiley and Sons.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater*^{SM} *v1.0* (TOEFL Research Report No. 62). ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02148.x

Appendix A. 30 Keywords

1. 安全 (safety)
2. 保护 (protection)
3. 超过 (exceed)
4. 乘坐 (ride)
5. 打扮 (dress up)
6. 发展 (development)
7. 放弃 (abandon)
8. 丰富 (rich)
9. 负责 (responsible)
10. 鼓励 (encourage)
11. 关键 (key)
12. 合格 (qualified)
13. 怀疑 (doubt)
14. 集合 (collection)
15. 坚持 (insist)
16. 节约 (save)
17. 紧张 (nervous)
18. 进行 (proceed)
19. 可怜 (poor)
20. 浪费 (waste)
21. 理想 (ideal)

22. 麻烦 (troublesome)

23. 缺点 (disadvantage)

24. 商量 (discuss)

25. 失望 (disappointing)

26. 无聊 (boring)

27. 详细 (detail)

28. 重视 (pay attention to)

29. 准时 (on time)

30. 尊重 (respect)

Appendix B. Samples of contextual AI-generated items (to select the most appropriate answer)

1 亲戚不时的批评让原本____的亲子关系更加恶劣 [单选题]

A放松

B紧张(Key)

C舒张

D严重

2 比赛的组织方邀请男篮教练教授当地农民篮球_____者打篮球 [单选题]

A 爱情

B 天分

C 厌恶

D爱好(Key)

Appendix C. Samples of expert-designed items

1. 他一考试就_____[单选题] *

A可惜

B紧张(Key)

C可怜

D可怕

2. 他每天认真讲课，是一个_____的老师[单选题] *

A合格(Key)

B活泼

C有用

D合适