The guessing from context test

Yosuke Sasao and Stuart Webb
Kyoto University, Japan | The University of Western Ontario, Canada

This study aims to develop two equivalent forms of the Guessing from Context Test (GCT) and provide its preliminary validity evidence. The GCT is a diagnostic test of the guessing skill and measures the following three important steps in guessing: identifying the part of speech of an unknown word, finding its discourse clue, and deriving its meaning. The test was administered to 428 Japanese learners of English. The results indicate that the two forms each with 20 question sets are equivalent in terms of item difficulty distribution and representativeness of the construct being measured. A wide range of validity evidence was provided using Messick's validation framework, the Rasch model, qualitative investigations into the relationships to actual guessing, and proposals for score interpretation.

Keywords: vocabulary, guessing from context, diagnostic test, equivalent forms, validation.

1. Introduction

The skill of guessing the meanings of unknown words from context plays an important part in vocabulary learning through reading and listening, because it is the most frequent and preferred strategy when learners deal with unknown words in context (Cooper, 1999; Fraser, 1999; Paribakht & Wesche, 1999). However, learners often fail in guessing. For example, Nassaji (2003) reported that the success rate was only 25.6% (44.2% even if partially correct guesses were included). Parry (1991) found that the success rate ranged from 12% to 33%. These low rates suggest a need for improvement in the guessing skill.

Although successful guesses do not always lead to learning (e.g., Brown, Waring, & Donkaewbua, 2008; Horst, Cobb, & Meara, 1998; Waring & Takaki, 2003), guessing makes a significant contribution to word retention. Guessing is a productive strategy that requires an active cognitive process including hypothesis testing about word meaning (Ellis, 1994; Haastrup, 1991). Meeting words in context provides a cognitive hook for word retention (Schouten-van Parreren, 1996). Guess-

https://doi.org/10.1075/itl.00009.sas
ITL - International Journal of Applied Linguistics 169:1 (2018), pp. 115–141. issn 0019-0829 | e-issn 1783-1490

ing word meanings followed by consulting a dictionary leads to a better retention of words (Fraser, 1999). It should be reasonable to assume that the improved skill of guessing has the potential to facilitate vocabulary learning, because it provides learners with a greater chance to learn words while reading or listening.

Despite the importance of the guessing skill, very few attempts have been made to develop a diagnostic test measuring learners' guessing skill. The guessing skill has been investigated using think-aloud protocols where learners verbalize what they think while guessing (e.g., Ames, 1966). This approach may have the advantage of providing learners with individualized diagnostic information, but the test administration and grading are demanding for teachers. This indicates a need for a test that is easy to administer and grade.

Another way of measuring the guessing skill is to use a multiple-choice format (Carnine, Kameenui, & Coyle, 1984; Nagy, Anderson, & Herman, 1987; Nagy, Herman, & Anderson, 1985; Schatz & Baldwin, 1986). One limitation to the existing tests is that they were developed for research purposes, and not for diagnostic purposes. They typically measure one aspect of guessing (deriving the meaning of unknown words), and as such, little diagnostic information is available on how the guessing skill may be improved. In addition, very few attempts have been made to validate the tests or create equivalent forms that allow a pre- and post-test design to measure the improvements of learners' guessing skill.

In order to fill this gap, the present research developed the Guessing from Context Test (GCT) that has the following characteristics:

1. It is easy to complete and grade;
2. It diagnoses learners' guessing skill;
3. It has two equivalent forms; and
4. Its preliminary validity evidence is provided.

The first two points are discussed in the subsequent section. The third and the final points are discussed in the *Development of two equivalent forms* and the *Test evaluation* sections, respectively.

2. Features of the GCT

2.1 What clues are included?

The GCT aims to provide learners with diagnostic information on their guessing skill. In so doing, it measures whether they can find and use clues in context. Among various types of clues (see, for example, de Bot, Paribakht, & Wesche, 1997; Haastrup, 1985, 1987, 1991; Nassaji, 2003, for classification of clues available

in guessing), it deals with grammar (part of speech of the unknown word) and discourse (relationships with other words or phrases in the context) clues. There are at least three reasons for the inclusion of these two types of clues in the GCT.

First, research has shown that the skills of using discourse clues (e.g., Fukkink & de Glopper, 1998; Kuhn & Stahl, 1998; Walters, 2006) and analyzing the grammatical structure in a sentence (e.g., Carpay, 1974; van Parreren, 1975) can be improved by teaching. These two types of knowledge are different from other clues such as L1 and world knowledge which may facilitate guessing but are often difficult to teach for teachers with different educational, professional, and L1 backgrounds from their students.

Second, although grammar and discourse clues may not always be helpful (Beck, McKeown, & McCaslin, 1983; Schatz & Baldwin, 1986), they are present in every context; that is, an unknown word always has a grammatical function in a sentence and is used in discourse. These clues are different from other clues such as morphological and world knowledge which are not always present.

Finally, searching for grammar and discourse clues is included in studies on practical procedures to help learners to successfully guess words from context (Bruton & Samuda, 1981; Clarke & Nation, 1980; Nation & Coady, 1988; Williams, 1985). Other types of clues are often regarded as a supportive strategy or excluded from the proposed steps. For example, Clarke and Nation (1980) excluded the use of background knowledge because it is not always available and is less likely to lead to vocabulary learning. In their study, the use of word part knowledge only plays a supportive role in checking the guess, because word part analysis is sometimes misleading[1] (Bensoussan & Laufer, 1984; Laufer & Sim, 1985; Nassaji, 2003).

Grammar clues are useful in deriving a general meaning of an unknown word. Clarke and Nation (1980, p. 212) argue that knowing the part of speech of a word allows the "Who does what to whom?" analysis. For example, in the sentence *Typhoon Vera killed or injured 218 people and crippled the seaport city of Keelung* (*crippled* is the target word to guess), learners may think that Typhoon Vera did something (=crippled) to Keelung because *crippled* is a verb. What a typhoon does is likely to have a negative influence on a city. This analysis may not be sufficient to arrive at the precise meaning of *cripple*, but together with the phrase *killed or injured 218 people*, learners may be able to guess its meaning as "damage" or "destroy." Clarke and Nation also emphasize the importance of grammar by argu-

1. Increasing morphological knowledge may contribute to improving the guessing skill, because morphological analysis is a frequently used strategy when guessing from context (de Bot, et al., 1997; Nassaji, 2003). Together with an affix test such as the Word Part Levels Test (Sasao & Webb, 2017), the GCT may provide more comprehensive information about guessing ability.

ing that failures in guessing seem to be frequently caused by misunderstanding the part of speech of the unknown word.

The GCT focuses on nouns, verbs, adjectives, and adverbs, because these four parts of speech account for the vast majority of word types in English. The ratio of target words for each part of speech was (noun): (verb): (adjective): (adverb) = 9:6:3:2 to reflect the British National Corpus (BNC) frequency data (Leech, Rayson, & Wilson, 2001).

The instruction of discourse clues may also be helpful, because even L1 high-school, undergraduate and graduate students are not always aware of a variety of discourse clues (McCullough, 1943; Strang, 1944). The GCT includes twelve types of discourse clues (direct description, indirect description, contrast/comparison, synonym, appositive, modification, restatement, cause/effect, words in series, reference, association, and example) that were identified by a total of nine studies: Six of them (Artley, 1943; Deighton, 1959; Dulin, 1970; Johnson & Pearson, 1984; Spache & Berg, 1955; Walters, 2006) relied on analysis of written texts for the classification of discourse clues, while the other three studies (Ames, 1966; McCullough, 1945; Seibert, 1945) classified discourse clues based on data from learners who guessed the meanings of words. It should be noted here that the taxonomies of discourse clues vary widely according to researchers. Some clues (e.g., direct description) are included in all the studies, while others (e.g., example) are not. Different researchers use different labels to refer to largely the same notion (e.g., *direct explanation* and *definition*), and the twelve discourse clues are not mutually exclusive.

2.2 How is the guessing skill measured?

The GCT has 20 sets of questions, each of which consists of a passage and three questions. The three questions individually measure the three important steps in guessing; that is, identifying the part of speech of an unknown word, finding its discourse clue, and guessing its meaning. Figure 1 illustrates a sample question set.

In the GCT, one target word is embedded in one passage. The passages were selected from the BNC and were paraphrased using the most frequent 1,000 word families in Nation's (2006) BNC word lists to the extent possible. Simplification was made to remove low-frequency words, and not to change the content or discourse clues. Each passage includes one of the twelve types of discourse clues selected for the GCT, and consists of 50–60 running words in order to achieve the 98% coverage which is considered to be desirable for successful guessing to occur (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006).

To reduce the likelihood that the target words are known, they were randomly chosen from the words included in the 11th to 14th 1,000 word families in Nation's

[Passage] Probably the world's finest (1)collection of 2,000-year-old cups will be shown at the (2)museum. From the 10th to 25th of October the show is held about various ways of having **duterages** such as (3)tea and coffee. Some of the cups on show are taken from the collection of an English man who gave them to the museum in 1979.

[Question 1] Choose the part of speech of the bold, underlined word.
(1) Noun
(2) Verb
(3) Adjective
(4) Adverb

[Question 2] Choose the word or phrase that helps you to work out the meaning of the bold, underlined word.
(1) collection
(2) museum
(3) tea and coffee

[Question 3] Guess the meaning of the bold, underlined word.
(1) food
(2) cup
(3) drink

Figure 1. Sample items

(2006) BNC word lists. The target words were replaced by nonsense words (words that do not exist in English) to ensure that the word forms were unknown to them. In the example, the original word was *beverages* which was replaced by the nonsense word *duterages*. The nonsense words had the same inflectional (e.g., *-ed* and *-s*) and derivational suffixes (e.g., *-ly* and *-ness*) as the original words to indicate their syntactic properties. In the example, the inflectional suffix *_s* was added to the nonsense word to indicate it is plural.

The order of the three questions (part of speech, discourse clue, and meaning) was determined based on Clarke and Nation's (1980) procedure for guessing. The first question asks about the part of speech of the target word. The second question aims to measure whether test-takers can find a discourse clue that helps guess its meaning. The correct answer is the word or phrase that includes one of the twelve discourse clues selected for the GCT. The distractors are of little use in deriving its meaning. In Figure 1, the correct answer is Option 3 where the target word's

examples are shown after the phrase *such as* (example clue). The third question measures whether test-takers can derive the meaning of the target word. Three options with the same part of speech are provided. The two distractors share some common meaning with the correct answer but contain irrelevant or lack important meaning. In the example, the correct answer is Option 3 *drink* which best fits to the context and is most similar in meaning to *beverage*. Options 1 (*food*) and 2 (*cup*) share a similar notion relating to eating or drinking, but they are not the best answers.

A potential weakness of this format is that it cannot measure the ability to use global clues which are found further away from the target word, because each passage needed to be relatively short (around 50 words) so that the test included a sufficient number of items to provide reliable results. However, immediate clues may be much more important than global ones, because in many cases learners arrive at successful guessing based on immediate rather than global clues, and poor guessers often have difficulty using immediate clues (Haynes, 1993; Morrison, 1996).

3. Development of two equivalent forms

3.1 Materials preparation

A series of pilot studies was conducted to ensure that (1) the simplified passages were comprehensible, (2) the discourse clues were identifiable, and (3) the target words were guessable. A small group of native English-speaking MA and PhD students individually read the passages, underlined the words that helped guess the meaning of the target word, and guessed its meaning without any options. The test was repeatedly piloted until no significant problem was found.

Next, multiple-choice questions were written and examined with another small group of native and non-native English speakers of high proficiency. The wrongly answered items were inspected and rewritten where necessary. Based on the piloting, a total of 60 sets of questions (5 passages × 12 discourse clues) were created.

The test was also piloted with ten Japanese learners of English with a wide range of proficiency levels to estimate the administration time. The instructions and the example items were also rewritten until they were readily comprehensible to them. The results indicated that they would need 1.5 minutes per question set.

3.2 Participants

A total of 428 Japanese learners of English as a foreign language participated in the research (277 males and 151 females; 221 high-school and 207 university students). The participants' ages ranged between 16 and 21 with the average being 17.7 (SD = 3.2). The high-school students had at least three years of prior English instruction, and the university students had been learning English for at least six years. Their majors included economics, engineering, law, literature, and pharmacology. The participants' self-reported TOEIC® scores from 134 students were: Mean = 425.2, SD = 182.2, Max = 910, and Min = 200.

3.3 Materials

The materials were created to develop an item difficulty scale for the 60 question sets and select good performing items based on the Rasch model (Rasch, 1960). The Rasch model was used for item analysis, because it produces an interval scale for item difficulty and person ability, provides fit statistics that help examine the degree of match between the observed data and the Rasch unidimensional model, and allows test equating where all items are put into one item hierarchy.

Six different forms each with 20 question sets were created, because 30 minutes (1.5 minutes × 20 sets) of test time was considered manageable for high-school students. As shown in Figure 2, the 60 question sets in the GCT were randomly classified into six groups (Item groups a-f) each with ten question sets. Six forms (Forms 1–6) were created by systematically combining the items in two of the six item groups. Each form consisted of a total of 20 items, ten of which overlapped with another form and the other ten of which overlapped with another different form. This systematic link was designed to conduct a concurrent (or one-step) equating where all the data are entered into one big array and the items that were not taken by a test-taker are treated as missing data. Although this design allows a large number of missing data, researchers (Bond & Fox, 2015; Linacre, 2016b) argue that Rasch analysis is robust with missing data which can be used intentionally by design. The test was written in a paper-based format. The information sheet, the consent form, and the instructions were translated into Japanese, the participants' L1.

3.4 Item analysis

Data were collected in October and November 2010. The six test forms were randomly distributed to the participants. Descriptive statistics for the six forms are summarized in Table 1.

Figure 2. Test design

Item group	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6
a (10 sets)	✔					✔
b (10 sets)	✔	✔				
c (10 sets)		✔	✔			
d (10 sets)			✔	✔		
e (10 sets)				✔	✔	
f (10 sets)					✔	✔

Table 1. Descriptive statistics for the six forms of the GCT

		Part of speech		Discourse clue		Meaning	
Form	**No. of participants**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	71	13.5	4.0	9.1	3.7	7.9	3.5
2	68	15.3	3.4	9.9	3.2	9.0	2.5
3	76	14.7	3.4	10.4	3.6	10.0	3.1
4	76	16.2	3.9	11.2	3.9	10.5	3.9
5	57	15.6	3.9	11.7	3.9	10.7	4.3
6	80	14.6	3.9	11.8	3.6	10.3	3.1
Total	**428**	**14.9**	**3.8**	**10.7**	**3.8**	**9.7**	**3.5**

Dichotomous Rasch analysis was performed for each question type using WINSTEPS 3.92.1 (Linacre, 2016a). Items were regarded as misfit if the point-measure correlation was smaller than .1, or the standardized fit statistics (outfit *t* and infit *t*) were larger than 2.[2] The results showed that 2, 5, and 4 non-overlapping items were identified as misfit for the part of speech, discourse clue, and meaning questions, respectively. The 11 question sets with misfit items were excluded from the GCT. The 49 acceptable question sets had 24 noun, 13 verb, 7 adjective, and 5 adverb target words. Three or more question sets survived for each of the twelve discourse clues.

2. The standardized fit statistics of smaller than −2 are called overfit. Overfit items do not indicate the same threat to the measurement quality as underfit (infit $t > 2$ or outfit $t > 2$) items. Overfit indicates that the data seem to show a Guttman pattern due to less variability than the model expectation. Each question type had less than 5% of the overfitting rate which is unlikely to affect item and person estimates substantially (Smith Jr., 2005). Thus, no treatment was made to the overfit items.

3.5 Creating equivalent forms

Equivalent forms are of the same test length, show the same item difficulty distribution, and are representative of the construct being measured. First, the test length was determined so that the estimated Rasch person strata index would be larger than 2, which indicates two statistically distinct levels for person abilities. A person strata of 2 is required for a test to be sensitive to gains from an experimental intervention such as teaching (Wolfe & Smith Jr., 2007). The person strata of 2 is equivalent to person reliability of .610 given the formulae in Linacre (2016b).[3] The number of items needed for achieving the reliability of .610 was estimated based on the Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910). Table 2 shows the estimated number of items that are required to arrive at the person strata of 2. The largest number of items (19.8) was estimated for the meaning question of Form 2. This indicates that a new test form should involve at least 20 items in order for any form to guarantee the minimum requirement for a sensitive test (Rasch person strata of 2).

Table 2. Estimated number of items needed for achieving person strata of 2

Question type	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6
Part of speech	9.3	14.0	16.0	6.6	14.5	16.7
Discourse clue	18.5	19.0	13.0	9.6	16.5	11.8
Meaning	13.1	19.8	11.9	11.6	8.9	12.9

Two equivalent 20-item test forms (Forms A and B) were created based on the following criteria to maintain the representativeness of the construct being measured:

1. Each form had 9 noun, 6 verb, 3 adjective, and 2 adverb target words in order to reflect actual language use.
2. Each form included all twelve types of discourse clues.
3. To ensure that each form has items with a wide spread of difficulty, the 49 acceptable items were classified into four groups based on the item difficulty estimates of the meaning items: (1) larger than 0.5 logits,[4] (2) between 0 and 0.5 logits, (3) between −0.5 and 0 logits, and (4) smaller than −0.5 logits. The item difficulty of the meaning question was used instead of the other questions, because deriving the meaning is arguably the most important aspect in guessing from context. Each form had five items selected from each of the four groups except for Form A with four items of the most difficult group and

3. Reliability = $G^2/(1+G^2)$, and Strata = $(4G+1)/3$, where G = separation coefficient.

six items of the second most difficult group, because there were only a total of nine items that showed difficulty estimates of larger than 0.5 logits.

The item distributions of the meaning question are shown in Figure 3 using a Rasch person-item map, which displays both persons in terms of ability and items in terms of difficulty on a Rasch interval scale. The far left of this figure shows a Rasch logit scale with the mean item difficulty being 0. This figure has two distributions on the logit scale: persons on the left and items on the right. More able persons and more difficult items are located towards the top, and less able persons and less difficult items are located towards the bottom. For the person distribution, each "#" represents three persons and each "." represents one or two persons. For the item distribution, the items of Form A are shown in the left and those of Form B on the right. Each number indicates the original item number followed by its Rasch item difficulty in brackets. The person and item distributions are interrelated in that a person has a 50% probability of succeeding on an item located at the same point on the logit scale. Figure 3 shows that there are few gaps in the item difficulty hierarchy and the item difficulties are largely evenly distributed between Forms A and B.

Levene's test was performed to examine the homogeneity of variance of the Rasch item difficulty estimates between the two forms. The null hypothesis of equal variances was not rejected at $\alpha = .05$ ($F = 2.18$, $p = .148$ for the part of speech; $F = 1.81$, $p = .187$ for the discourse clue; and $F = 0.00$, $p = .957$ for the meaning questions), indicating that the spread of item difficulties may be acceptably equal between the two forms. Subsequent *t*-tests (2 tailed) did not detect any significant differences in the mean item difficulties between the two forms for any of the three sections (Table 3). The effect sizes (*r*) were smaller than .2 which indicates small differences between the two forms (Cohen, 1988, 1992). Taken together, the two forms may be equivalent in terms of difficulty as well as representative of the construct being measured.

Table 3. Comparison of the item difficulties between the two forms

	Form A		Form B					
	M	*SD*	*M*	*SD*	*t*	*d.f.*	*p*	*r*
Part of speech	−0.06	1.12	0.01	1.41	−0.17	38	.866	.027
Discourse clue	−0.11	0.46	0.03	0.69	−0.78	38	.440	.119
Meaning	0.07	0.73	0.07	0.74	−0.01	38	.995	.001

4. Logit is the contraction of log-odds unit (of success), the unit of measurement of item difficulty and person ability estimates. Larger logit values indicate more difficult items or more able persons, and *vice versa*.

Figure 3. Person-item map of the equivalent forms for the meaning question

<More able persons>			<More difficult items>						
				Form A			Form B		
	*##								
2	##		+						
	*#	T		40(1.65)					
	*#			T			34(1.34)		
	###			14(1.29)			30(1.16)	28(1.15)	
1	#####	S	+	42(0.94)			7(0.98)	39(0.96)	
	*######			S 13(0.76)					
	########			24(0.48)	26(0.36)		1(0.38)		
	###############			32(0.32)	2(0.31)	38(0.23)	15(0.22)	33(0.20)	
0	##############	M	+	M 17(0.14)	10(−0.12)	56(−0.13)	53(0.11)	5(0.04)	8(−0.02)
	*############			23(−0.16)	35(−0.26)		25(−0.18)	50(−0.18)	18(−0.25)
	##########			6(−0.32)			46(−0.47)	36(−0.57)	57(−0.62)
	*#########			S 45(−0.68)	59(−0.78)	55(−0.81)	20(−0.63)		
−1	**#######	S	+	27(−0.87)	48(−1.00)		12(−1.04)		
	####						44(−1.20)		
	#			T					
	*##	T							
−2	*		+						
	*								
<Less able persons>			<Less difficult items>						

4. Test evaluation

This section aims to provide preliminary validity evidence for the GCT from five aspects of construct validity (content, substantive, structural, generalizability, and external aspects) proposed by Messick (1989, 1995). It also reports on a small-scale qualitative investigation into the relationship with actual guessing, and discusses ways in which the scores are interpreted and presented to learners.

4.1 Content aspect of construct validity

This aspect aims to clarify "the boundaries of the construct domain to be assessed" (Messick, 1995, p. 745). It addresses the relevance, representativeness and technical quality of the items. The content relevance to the guessing skill was discussed in the *Features of the GCT* section.

The GCT is considered to be representative of the construct domain, because (1) the target words were randomly selected from low-frequency words, (2) the ratio of the four parts of speech reflects authentic language use, (3) a wide variety of discourse clues are included, and (4) there are few gaps in the item difficulty hierarchy (Figure 3). Representativeness may also be evaluated by Rasch item strata which indicate the number of statistically different levels of item difficulty. The item strata statistics were above 2 (6.07, 3.57, and 4.85 for the part of speech, discourse clue and meaning questions, respectively) which is the minimum requirement for interpretable scores (Smith Jr., 2004, p. 106).

Technical quality may be examined by Rasch item fit statistics (Smith Jr., 2004). No question sets with any misfit items were used for the new forms, which indicates a high degree of technical quality of the new test forms.

4.2 Substantive aspect of construct validity

This aspect refers to "theoretical rationales for the observed consistencies in test responses […] along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (Messick, 1995, p. 745). It was difficult to predict a single factor that significantly affected the item hierarchy as shown in Figure 3, because guessing is a complex cognitive process (de Bot, et al., 1997; Haastrup, 1987, 1991; Nassaji, 2003). It was hypothesised that the success in the meaning items would depend more on knowledge of discourse clues than that of part of speech. As discussed earlier, knowledge of part of speech may help derive a partial meaning such as positive/negative or person/thing, but in many cases, discourse clues are necessary for deriving a precise meaning. Nassaji (2003) found that the use of discourse clues contributed to more successful

guesses (55.6%) than the use of grammatical clues (41.7%). It was also hypothesized that a combination of the part of speech and discourse clue knowledge would make a significant contribution to successful guessing, because these types of knowledge play an important role in guessing (Bruton & Samuda, 1981; Clarke & Nation, 1980; Williams, 1985).

These two hypotheses were examined using a multiple regression analysis, where the dependent variable was the person ability estimates from the meaning items and the independent variables were those from the part of speech and the discourse clue items. Figure 4 presents a path diagram of the multiple regression analysis (without correction for attenuation due to measurement error).[5] This figure shows that the β coefficient for the discourse clue items (.44) was higher than that for the part of speech items (.32). In addition, a combination of the part of speech and discourse clue knowledge accounted for about half of the variability of the ability to derive the meaning ($R^2 = .45$). Given that guessing involves many other factors such as reading ability and world knowledge, this coefficient of determination may be considered high. Taken together, the observed data seem to be consistent with the theoretical rationales.

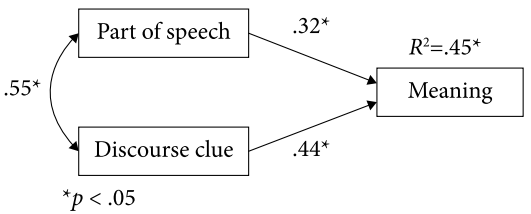

Figure 4. Relationships of the part of speech and the discourse clue items to the meaning items

The second hypothesis was also tested by looking at the scores of good guessers. Out of the 428 participants, 48 students were found to be at an advanced level for the meaning question (Rasch person ability of greater than 1; see the *Score interpretation* section for the classification of the guessing skill). Table 4 presents their scores of the part of speech and the discourse clue questions. This table shows the general tendency that the students skillful at deriving word meanings are also good at identifying the part of speech and the discourse clues. No participants were at a beginner level for the two questions. Two students marked an intermediate level for the part of speech and/or the discourse clue questions,

5. No serious sign of multi-collinearity was detected. The variance inflation factor (VIF) was 1.45 for both the part of speech and the discourse clue questions, which is below 10 (threshold level for multi-collinearity).

because they left about half of the items unanswered. This may indicate that knowledge of both part of speech and discourse clues is needed for successful guessing.

Table 4. Score distribution of the participants at an advanced level for the meaning question

Guessing skill level		
Part of speech	**Discourse clue**	**No. of participants**
Advanced	Advanced	34
Advanced	Upper-intermediate	9
Advanced	Intermediate	1
Upper-intermediate	Advanced	2
Upper-intermediate	Upper-intermediate	1
Intermediate	Intermediate	1
Total		**48**

The substantive aspect of construct validity was also evaluated by examining Rasch person fit. As with item fit, a misfit person was defined as outfit $t > 2$ or infit $t > 2$ (underfit), or outfit $t < -2$ or infit $t < -2$ (overfit). Each question type had the misfit rate of less than 5% which was expected to occur by chance given the nature of the z distribution. This indicates that the test-takers' response pattern corresponded to the modelled difficulty order.

4.3 Structural aspect of construct validity

This aspect "appraises the fidelity of the scoring structure to the structure of the construct domain at issue" (Messick, 1995, p. 745). It includes the evaluation of unidimensionality (the degree to which a test measures one attribute at a time), because a unidimensional measure allows straightforward scoring. Linacre (1995) suggested that dimensionality may be addressed by (1) item correlations, (2) fit statistics, and (3) principal components analysis (PCA) of standardized residuals. The new forms were created from the items without any problems in terms of item correlation and fit. The PCA of standardized residuals was performed, and the scree plot for each question type is presented in Figures 5–7. These figures show that the eigenvalues of the first contrast (largest secondary dimension) are 2 or less which may occur by chance (Linacre & Tennant, 2009; Raîche, 2005), and the eigenvalues of other contrasts seem to reach an asymptote at the first contrast (see Stevens, 2002; Wolfe & Smith Jr., 2007 for a detailed discussion). This may be taken as positive evidence for unidimensionality of the GCT.

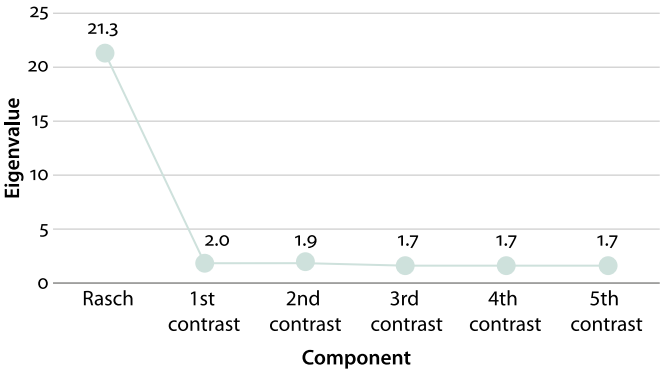

Figure 5. Scree plot for the part of speech items

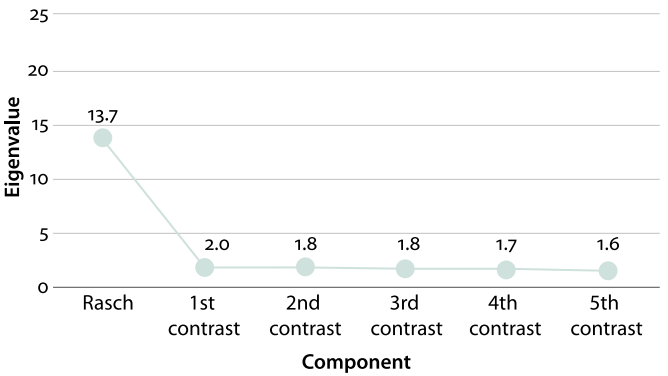

Figure 6. Scree plot for the discourse clue items

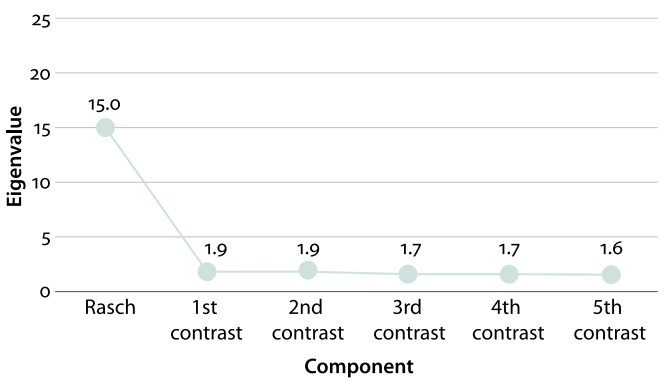

Figure 7. Scree plot for the meaning items

4.4 Generalizability aspect of construct validity

This aspect deals with "the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks" (Messick, 1995, p. 745). It was evaluated by examining the extent to which Rasch person ability and item difficulty estimates were invariant within the measurement error (Andrich, 1988; Smith Jr., 2004; Wolfe & Smith Jr., 2007; Wright & Stone, 1979).

Person measure invariance was examined by test reliability, or reproducibility of person ability measures. Table 5 shows Rasch person reliability (equivalent to traditional reliability coefficients such as Cronbach's alpha), and Rasch person separation which is linear and ranges from zero to infinite. The results showed that the reliability estimates ranged between .56 and .78 with the average being .66. The small number of items after the deletion of the misfit items may have affected the reliability (Linacre, 2016b), but the average reliability of .66 does not seem unacceptably low. Fukkink and de Glopper (1998) conducted a meta-analysis of twelve previous studies on the effects of teaching on the guessing skill, and reported that the tests used in these studies had the average Cronbach's alpha of .63 (Max = .85, Min = .13). The low reliability estimates may be understandable, because the construct of guessing from context is complex including a wide range of language ability.

Table 5. Rasch person separation and reliability

			Part of speech		Discourse clue		Meaning	
	No. of items	No. of participants	PS	PR	PS	PR	PS	PR
Form 1	17	71	1.67	.74	1.21	.59	1.44	.67
Form 2	19	68	1.47	.68	1.47	.68	1.24	.60
Form 3	13	76	1.12	.56	1.25	.61	1.32	.63
Form 4	15	76	1.87	.78	1.58	.71	1.41	.67
Form 5	18	57	1.39	.66	1.39	.66	1.79	.76
Form 6	16	80	1.23	.60	1.47	.68	1.39	.66

* PS = person separation; PR = person reliability

Person measure invariance was also examined by dividing the items into the first and the second halves and conducting DPF (Differential Person Functioning) analysis to examine a practice or fatigue effect. No statistically significant DPF was detected for any persons for the three question types ($\alpha = .05$). In other words, no practice or fatigue effect was observed statistically.

Item calibration invariance was examined by analyzing Rasch item reliability, which has no traditional equivalent and addresses the degree to which item

difficulties are reproducible. Table 6 shows that the reliability estimates ranged between .69 and .94 with the average being .84. This indicates that the item difficulty estimates are highly reproducible.

Table 6. Rasch item separation and reliability

	Part of speech		Discourse clue		Meaning	
	IS	IR	IS	IR	IS	IR
Form 1	3.80	.94	1.50	.69	1.83	.77
Form 2	2.71	.88	2.17	.82	2.37	.85
Form 3	2.90	.89	1.77	.76	2.18	.83
Form 4	3.65	.93	2.19	.83	1.80	.76
Form 5	2.74	.88	2.20	.83	2.51	.86
Form 6	2.78	.89	2.71	.88	3.14	.91

* IS = item separation; IR = item reliability

Next, the DIF (Differential Item Functioning) analysis was performed to examine whether the item calibrations from male ($N=277$) and female ($N=151$) test-takers are invariant for each of the three question types. Welch's t-test revealed that statistically significant DIF was detected for one item in each question type ($\alpha=.05$). A qualitative inspection of these items did not find any reason for this DIF. The DIF rate is 2.5% (1/40 items per question type), which is less than 5% which may occur by chance given the nature of Type I error.

4.5 External aspect of construct validity

This aspect refers to "the extent to which the test's relationships with other tests and nontest behaviors reflect the expected high, low, and interactive relations implied in the theory of the construct being assessed" (Messick, 1989, p. 45). The relationships between the GCT scores (Rasch person ability estimates) and self-reported TOEIC scores were examined. It was hypothesized that the TOEIC and GCT scores would be positively correlated, because TOEIC is a test of English reading and listening skills which may involve the skill of guessing from context as an important component. However, it was also hypothesized that the GCT-TOEIC correlations would be lower than the within-GCT correlations (those between the scores from any two questions of the GCT), because the three question types in the GCT measure different aspects of the guessing skill. Table 7 presents a matrix of the Pearson's product-moment correlation coefficients between the GCT and TOEIC scores.[6] It shows that the GCT and TOEIC scores correlated

positively (r = .239, .295, .463), but the GCT-TOEIC correlations were lower than the within-GCT correlations (r = .550, .608, .658).

Table 7. Correlations between GCT and TOEIC scores

	Part of speech	Discourse clue	Meaning
Discourse clue	.550*		
Meaning	.608*	.658*	
TOEIC	.239*	.295*	.463*

N = 134;* p < .05.

In order to determine whether there are statistically significant differences between these two groups of correlation coefficients (GCT-TOEIC vs. within-GCT correlations), a Z-test was performed by means of a Meng-Rosenthal-Rubin method (Meng, Rosenthal, & Rubin, 1992). Table 8 shows that for all three question types, the within-GCT correlations were significantly higher than the GCT-TOEIC correlations (α = .05). This indicates that the above-mentioned hypothesis (positive but lower correlations for the GCT-TOEIC scores than the within-GCT correlations) may be acceptable.

Table 8. Difference between within-GCT and GCT-TOEIC correlations

Question type	within-GCT correlations	GCT-TOEIC correlations	Z	p
Part of speech	r_{PD} = .550	r_{PT} = .239	3.40	.001
	r_{PM} = .608	r_{PT} = .239	4.70	.000
Discourse clue	r_{DP} = .550	r_{DT} = .295	2.75	.006
	r_{DM} = .658	r_{DT} = .295	4.84	.000
Meaning	r_{MP} = .608	r_{MT} = .463	2.18	.029
	r_{MD} = .658	r_{MT} = .463	2.51	.012

Note. N = 134, P = part of speech, D = Discourse clue, M = meaning, T = TOEIC (e.g., r_{PD} = correlation coefficient between the part of speech and the discourse clue scores).

6. The TOEIC scores were provided by 134 out of the 428 (31.3%) of the participants. Welch's t-test did not find statistically significant difference between the Rasch person ability estimates of the 134 TOEIC-score reporters and the other 294 non-reporters (α = .05). The effect sizes (r) were small at .054, .089, and .125 for the part of speech, discourse clue, and the meaning items, respectively. This indicates that the results from the 134 reporters may be generalizable to the overall 428 participants.

4.6 Qualitative investigation

To examine the relationship with actual guessing, a recall version of the GCT (writing answers without any choices) was administered to a total of 14 English- or Japanese-native speaking graduate or undergraduate students in Japan with a variety of proficiency levels. They individually took the recall version with 30 randomly selected question sets and then the original version with the same 30 sets. For the recall version, they were asked to write answers in English or Japanese for the part of speech and the meaning questions and to underline a word or phrase for the discourse clue question.

The responses in the recall version were scored by a native English speaker with a high proficiency in Japanese and one of the authors (a native Japanese speaker). Inter-rater reliability was high (Spearman's $\rho = 1.00$ for the part of speech, .97 for the discourse clue, and .96 for the meaning questions). For the recall version, average raw scores from the two raters were used for analysis. Spearman's ρ between the original and the recall GCT scores were .91*, .77*, and .81* ($^*p < .05$) for the part of speech, discourse clue, and the meaning questions, respectively. A relatively low correlation (.77) was found in the discourse clue question, because one participant left nine items unanswered in the original version. If this person was excluded from the analysis, Spearman's ρ increased to .89. Taken together, the original and the recall versions of the GCT are strongly related and the constructs being measured overlap to a large extent.

4.7 Score interpretation

The item strata presented in the *Content aspect of construct validity* section (6.07, 3.57, and 4.85 for the part of speech, discourse clue, meaning questions, respectively) indicate that having three cut points (four levels) may be statistically justified. The discourse clue question showed the item strata index of smaller than 4, but it approached 4 and different cut points for different questions may make the score interpretation complicated. The three cut points were set at 1, 0, and −1 logits to create four levels, because the item difficulty estimates range from around −1 to 1 logits (Figure 3). The four guessing skill levels are summarized in Table 9. For easier interpretation, the corresponding raw scores are also provided as a rough approximation.

These four guessing skill levels were examined based on the TOEIC scores. Figures 8–10 illustrate the relationships between the self-reported TOEIC scores and the four levels for the part of speech, discourse clue, and meaning questions, respectively. The horizontal axis indicates the TOEIC score range, and the vertical axis shows the percentage of participants at each level as measured by the GCT.

Table 9. Guessing skill levels

		Raw score range		
Level	**Rasch ability range**	**Part of speech**	**Discourse clue**	**Meaning**
Advanced	Above 1 logits	16–20	16–20	16–20
Upper-intermediate	0 ~ 1 logits	13–15	11–15	11–15
Intermediate	−1 ~ 0 logits	10–12	6–10	6–10
Beginner	Below −1 logits	0–9	0–5	0–5

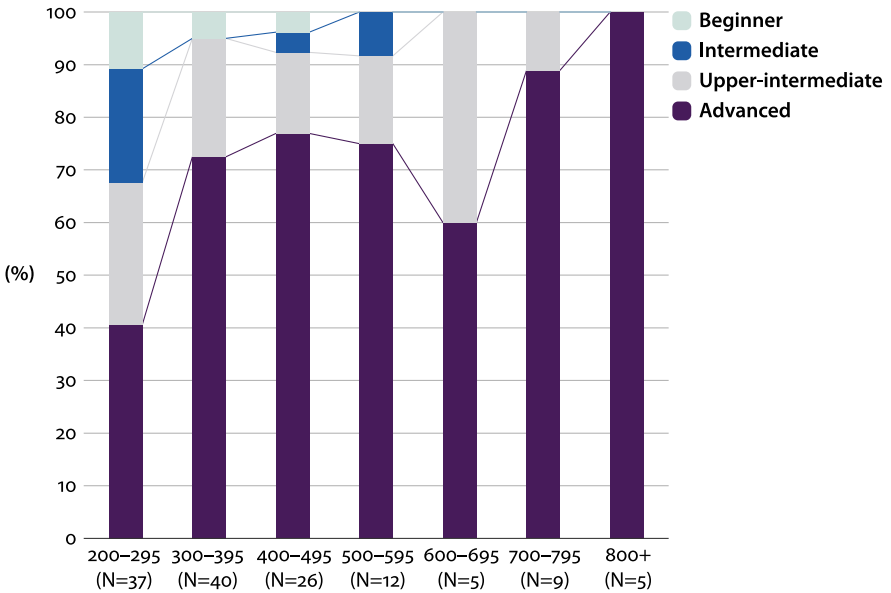

Figure 8. The four guessing skill levels for the part of speech question according to the TOEIC scores

These figures show the general tendency that the three components of the guessing skill improve as the general language proficiency develops. Figure 8 reveals that part of speech knowledge may be learned at an early stage (TOEIC scores of 300–395), but about a quarter of students with the TOEIC scores below 700 did not display their mastery of part of speech. This may indicate a need for the part of speech question which may help make them aware of the value of part of speech in guessing from context. It seems more difficult to reach the advanced level for the discourse clue than for the part of speech (Figure 9). For the meaning question, there is a marked increase in the percentage of students at the advanced level between the TOEIC scores of 600s and 700s (Figure 10). This indicates that the

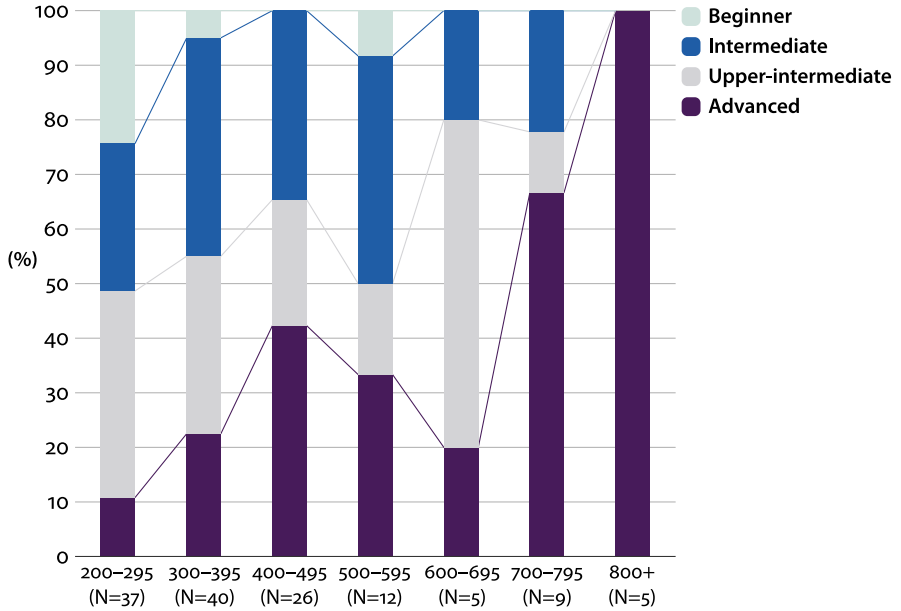

Figure 9. The four guessing skill levels for the discourse clue question according to the TOEIC scores

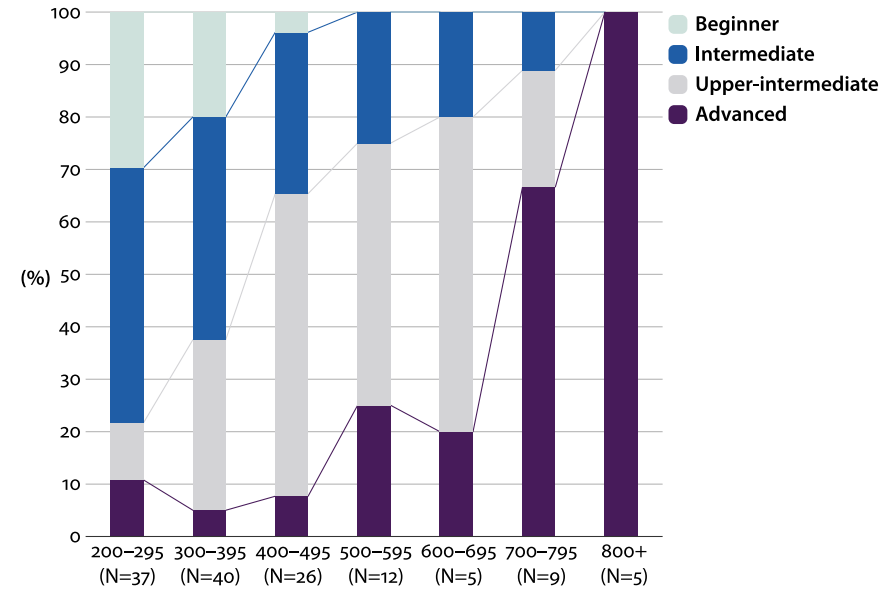

Figure 10. The four guessing skill levels for the meaning question according to the TOEIC scores

three components of the GCT may be useful in diagnosing the guessing ability of learners with a wide variety of proficiency levels.

For practical use of the GCT, diagnostic feedback needs to be easy to understand and clearly reveal learners' guessing skill. To meet this need, a bar graph may be useful because the information is visually presented and intuitively interpretable. Figure 11 shows a sample score summary of Learner A who got 19, 8, and 6 items correct for the part of speech, discourse clue, and meaning questions, respectively, with reference to Table 9 for the conversion between the raw scores and the corresponding levels. This graph shows that this learner demonstrated good knowledge of part of speech, but his weakness lies in finding discourse clues and deriving the meaning based on that information. This learner should focus on the learning of discourse clues to potentially improve guessing.

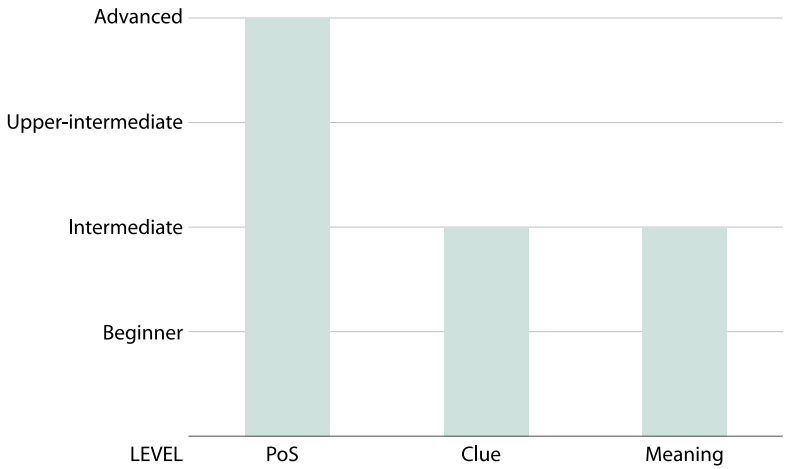

Figure 11. Score report example

5. Conclusion

The GCT was created to provide diagnostic information on learners' skill for guessing from context. It consists of 20 sets of a short (around 50 words) passage with three questions: part of speech, discourse clue, and meaning. Two equivalent forms were created to examine the effects of teaching and learning tasks. Both forms have 20 question sets to ensure that person item strata will be greater than 2. Preliminary validity evidence was provided for the GCT. The results generally indicated that the GCT is a reliable and valid measure of the guessing skill. For easy score interpretation, raw scores may be used to determine learners' levels of the guessing skill and indicate their weaknesses.

It should be noted that even if people do well on the GCT, this does not necessarily mean that they guess in the same way in real life situations as they do on the test. The GCT consists of texts with high-frequency words and clues to the target words are provided. In authentic texts, clues may only be available in unfamiliar words or may not be useful enough to derive the precise meaning (Laufer, 1997; Schatz & Baldwin, 1986).

It would be useful for future research to investigate effective teaching methods to improve the guessing skill. Different tasks and teaching materials may result in the development of different aspects of the guessing skill. A learner's proficiency level may also be an important factor affecting the effectiveness of instruction. Less proficient learners may benefit from general strategy instruction, while more advanced learners may need to be aware of specific types of discourse clues (Walters, 2006).

Future work will also involve the development of a web-based GCT so that the test can be administered easily and the feedback may be provided promptly. At the moment, the GCT is freely available in a paper-based format together with an answer key, and a summary of discourse clues at: http://ysasaojp.info/testen.html (February, 2017).

Acknowledgements

This research was supported by a Faculty Research Grant from Victoria University of Wellington, New Zealand (Grant ID: 110915), and JSPS KAKENHI Grant Number JP26770190.

References

Ames, W. S. (1966). The development of a classification scheme of contextual aids. *Reading Research Quarterly*, 2(1), 57–82. https://doi.org/10.2307/747039

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills, CA: Sage.

Artley, A. S. (1943). Teaching word-meaning through context. *Elementary English Review*, 20(1), 68–74.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *Elementary School Journal*, 83(3), 177–181. https://doi.org/10.1086/461307

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7(1), 15–32. https://doi.org/10.1111/j.1467-9817.1984.tb00252.x

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, 20(2), 136–163.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.

Bruton, A., & Samuda, V. (1981). Guessing words. *Modern English Teacher*, 8(3), 18–21.

Carnine, D., Kameenui, E. J., & Coyle, G. (1984). Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly*, 19(2), 188–204. https://doi.org/10.2307/747362

Carpay, J. A. M. (1974). Foreign-language teaching and meaningful learning: A Soviet Russian point of view. *ITL*, 25–26, 161–187.

Clarke, D. F., & Nation, I. S. P. (1980). Guessing the meanings of words from context: Strategy and techniques. *System*, 8(3), 211–220. https://doi.org/10.1016/0346-251X(80)90003-2

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly*, 33(2), 233–262. https://doi.org/10.2307/3587719

de Bot, K., Paribakht, T. S., & Wesche, M. (1997). Towards a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading. *Studies in Second Language Acquisition*, 19(3), 309–329. https://doi.org/10.1017/S0272263197003021

Deighton, L. C. (1959). *Vocabulary development in the classroom*. New York, NY: Columbia University Press.

Dulin, K. L. (1970). Using context clues in word recognition and comprehension. *Reading Teacher*, 23(5), 440–445.

Ellis, R. (1994). Factors in the incidental acquisition of second language vocabulary from oral input: A review essay. *Applied Language Learning*, 5(1), 1–32.

Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21(2), 225–241. https://doi.org/10.1017/S0272263199002041

Fukkink, R. G., & de Glopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, 68(4), 450–469. https://doi.org/10.3102/00346543068004450

Haastrup, K. (1985). Lexical inferencing – a study of procedures in reception. *Scandinavian Working Papers on Bilingualism*, 5, 63–87.

Haastrup, K. (1987). Using thinking aloud and retrospection to uncover learners' lexical inferencing procedures. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 197–212). Clevedon: Multilingual Matters.

Haastrup, K. (1991). *Lexical inferencing procedures or talking about words*. Tubingen: Gunter Narr.

Haynes, M. (1993). Patterns and perils of guessing in second language reading. In T. Huckin, M. Haynes & J. Coady (Eds.), *Second Language Reading and Vocabulary* (pp. 46–64). Norwood, NJ: Ablex.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11(2), 207–223.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.

Johnson, D., & Pearson, P. D. (1984). *Teaching reading vocabulary*. New York, NY: Holt, Rinehart & Winston.

Kuhn, M. R., & Stahl, S. A. (1998). Teaching children to learn word meanings from context. *Journal of Literacy Research*, 30(1), 119–138. https://doi.org/10.1080/10862969809547983

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 20–34). Cambridge: Cambridge University Press.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.

Laufer, B., & Sim, D. D. (1985). Taking the easy way out: Non-use and misuse of clues in EFL reading. *English Teaching Forum*, 23(2), 7–10, 20.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow: Longman.

Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions*, 9(2), 422–423.

Linacre, J. M. (2016a). WINSTEPS® Rasch measurement computer program. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2016b). *WINSTEPS® Rasch measurement computer programs User's Guide*. Beaverton, OR: Winsteps.com.

Linacre, J. M., & Tennant, A. (2009). More about critical eigenvalue sizes (variences) in standardized-residual principal components analysis (PCA). *Rasch Measurement Transactions*, 23(3), 1228.

McCullough, C. M. (1943). Learning to use context clues. *Elementary English Review*, 20, 140–143.

McCullough, C. M. (1945). The recognition of context clues in reading. *Elementary English Review*, 22(1), 1–5.

Meng, X. -L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1), 172–175. https://doi.org/10.1037/0033-2909.111.1.172

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morrison, L. (1996). Talking about words: A study of French as a second language learners' lexical inferencing procedures. *Canadian Modern Language Review*, 53(1), 41–75.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, 24(2), 237–270. https://doi.org/10.3102/00028312024002237

Nagy, W. E., Herman, P., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–253. https://doi.org/10.2307/747758

Nassaji, H. (2003). L2 vocabulary learning from context: strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly*, 37(4), 645–670. https://doi.org/10.2307/3588216

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P., & Coady, J. (1988). Vocabulary and reading. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 97–110). London: Longman.

Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition*, 21(2), 195–224. https://doi.org/10.1017/S027226319900203X

Parry, K. (1991). Building a vocabulary through academic reading. *TESOL Quarterly*, 25(4), 629–653. https://doi.org/10.2307/3587080

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, 19(1), 1012.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research*, 21(1), 12–30.

Schatz, E. K., & Baldwin, R. S. (1986). Context clues are unreliable predictors of word meaning. *Reading Research Quarterly*, 21(4), 439–453. https://doi.org/10.2307/747615

Schouten-van Parreren, C. (1996). Vocabulary learning and metacognition. In K. Sajavaara & C. Fairweather (Eds.), *Approaches to second language acquisition* (pp. 63–69). Jyvaskyla: University of Jyvaskyla.

Seibert, L. C. (1945). A study on the practice of guessing word meanings from a context. *Modern Language Journal*, 29(4), 296–323. https://doi.org/10.1111/j.1540-4781.1945.tb00276.x

Smith Jr., E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 93–122). Maple Grove, MN: JAM Press.

Smith Jr., E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, 6, 147–163.

Spache, G., & Berg, P. (1955). *The art of efficient reading*. New York, NY: Macmillan.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Strang, R. M. (1944). How students attack unfamiliar words. *The English Journal*, 33(2), 88–93. https://doi.org/10.2307/806504

van Parreren, C. F. (1975). First and second-language learning compared. In A. J. van Essen & J. P. Menting (Eds.), *The context of foreign-language learning* (pp. 100–116). Assen: Van Gorcum.

Walters, J. (2006). Methods of teaching inferring meaning from context. *RELC Journal*, 37(2), 176–190. https://doi.org/10.1177/0033688206067427

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15(2), 130–163.

Williams, R. (1985). Teaching vocabulary recognition strategies in ESP reading. *ESP Journal*, 4(2), 121–131. https://doi.org/10.1016/0272-2380(85)90015-0

Wolfe, E. W., & Smith Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 2 – Validation activities. *Journal of Applied Measurement*, 8, 204–234.
Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Address for correspondence

Yosuke Sasao
Kyoto University
Yoshida Nihonmatsu-cho, Sakyo-ku
Kyoto, 606-8501
Japan
sasao.yosuke.8n@kyoto-u.ac.jp