

Perspectives

A Contextualized Meaning-Recall Vocabulary Testing Platform

Tim Stoeckel

University of Niigata Prefecture

Stuart McLean

Kindai University

Paul Raine

Ritsumeikan University

Hung Tan Ha

University of Economics Ho Chi Minh City and Victoria

University of Wellington

Nam Thi Phuong Ho

University of Economics Ho Chi Minh City

Young Ae Kim

Kindai University and Kyoto Seika University

In contextualized vocabulary assessment, target words appear in extended context. Compared to tests employing single-word or limited-context prompts, research suggests that contextualized assessment is more reliable and demonstrates better concurrent validity. In meaning-recall vocabulary assessment, examinees retrieve

<https://doi.org/10.37546/JALTJJ45.2-2>

JALT Journal, Vol. 45, No. 2, November 2023

target-word meaning from memory and typically demonstrate knowledge via a written L2-to-L1 translation. Compared to multiple-choice formats, meaning-recall yields more reliable data, correlates more strongly with reading comprehension, and is less influenced by guessing and test strategies. To facilitate these approaches to vocabulary assessment, this article introduces a resource for teachers and researchers to create, administer, and mark contextualized meaning-recall tests. Users input a passage, select target items, and share the test URL with examinees. Examinees then provide L1 translations or L2 synonyms, definitions, or explanations of target words in input boxes below the lines of text. Raters mark responses online, and these judgments can be saved for partial automatic marking in future test use.

文脈化された語彙測定では、ターゲット項目が段落の中に現れる。単一語彙または限られた文脈の項目を用いるテストに比べて、文脈化された測定はより信頼性が高く、より優れた併存的妥当性を示している。意味想起語彙テストでは、受験者はターゲット語彙の意味を思い出し、通常は第二言語(L2)から第一言語(L1)への書記による翻訳で知識を示す。多肢選択形式と比較して、意味想起はより信頼性のあるデータをもたらし、読解力とより強い相関を示し、推測やテスト戦略の影響を受けにくい。語彙測定へのこれらのアプローチを支持するために、本稿では教師や研究者が文脈化された意味想起テストを作成、実施、採点するためのリソースを紹介している。利用者は文章を入力し、問題項目を選択、テストのURLを受験者と共有する。受験者は、テキストの下にある入力ボックスに対象語のL1翻訳やL2同義語、定義、またはその説明を入力する。採点者はオンラインで回答を評価し、正答とみなされる解答は将来のテスト利用時の自動採点のために保存できる。

Keywords: contextualized meaning-recall test; meaning-recall; vocabulary assessment

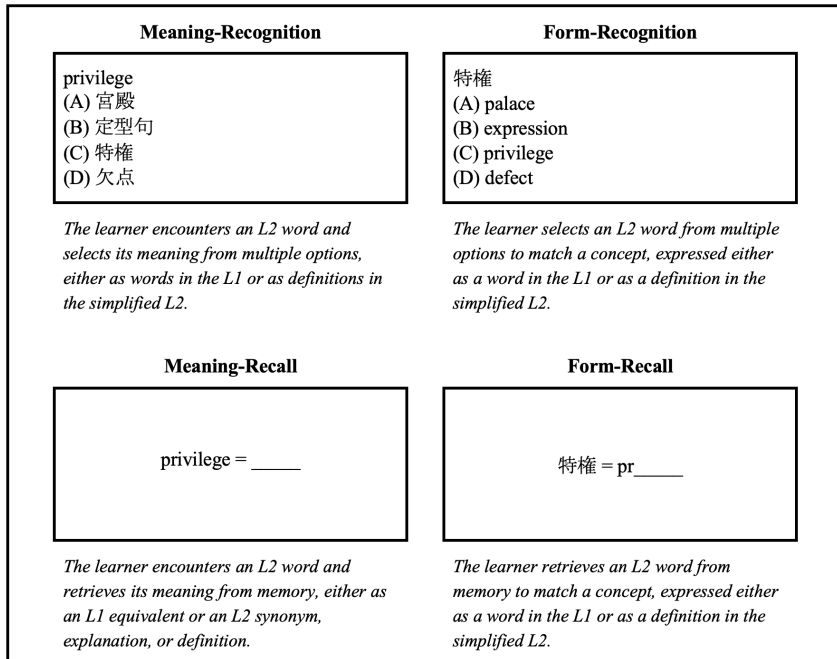
The availability of second language (L2) vocabulary assessment instruments of both breadth and depth has increased markedly over the past 20 years. One aspect of vocabulary knowledge commonly assessed with these tests is the form-meaning link, or the ability to associate meaning with the written or spoken form of a word (Jiang, 2002; McLean, Kramer, & Beglar, 2015). One kind of form-meaning assessment that has attracted recent attention is that of meaning-recall knowledge, or the ability to retrieve the meaning of an L2 word from memory upon seeing or hearing the word form. Meaning-recall is typically assessed by eliciting L2-to-L1 translations, or L2 synonyms or explanations of tested words. Meaning-recall tests are better predictors of reading ability than multiple-choice instruments (i.e., meaning-recognition; Stewart et al., 2023; Zhang & Zhang, 2022), making them attractive for many research purposes (Stewart et al., 2021; Stoeckel et al., 2021).

A drawback of meaning-recall assessment, it is sometimes argued, is that it is less practical, requiring more time to administer and mark tests (Webb, 2021a). One testing site, Vocableveltest.org (McLean et al., 2021), addresses

this problem with automated marking based on meticulously created banks of possible correct responses. Though this does not eliminate hand-marking, as novel responses do occur, it greatly reduces the time needed for scoring. A limitation of this tool, however, is that test makers must select from pre-existing lists of target words and test items. Though [Vocabularytest.org](https://vocabularytest.org) offers an extensive array of choices, teachers and researchers may at times wish to assess other words and, more importantly, in contexts other than those appearing in the existing item bank. To address this shortcoming, the present paper introduces a contextualized meaning-recall test (CMRT) platform designed to complement the assessment options offered by [Vocabularytest.org](https://vocabularytest.org). It differs from [Vocabularytest.org](https://vocabularytest.org) in that test makers input their own items and tests, meaning they can assess any target word or phrase desired. Moreover, vocabulary can be tested in contexts ranging from a single word to full-length passages. Though not as robust as the automated marking feature of [Vocabularytest.org](https://vocabularytest.org), the CMRT platform also allows for partial automated marking by saving manual ratings of responses for future test use. A beta version of the platform is currently available at <https://cmrt.vocabularytest.org/>. Though it shares a partial domain name with [Vocabularytest.org](https://vocabularytest.org), the two resources operate independently.

A Framework for Contextualized Assessment of Form-Meaning Knowledge

We begin by reviewing contextualized assessment of form-meaning knowledge. In form-meaning tests, vocabulary knowledge can be assessed at the levels of form-recognition, form-recall, meaning-recognition, and meaning-recall (Schmitt, 2010; Figure 1). The difference between recall and recognition is that in recall, examinees retrieve word meaning or form from memory, while in recognition they select meaning or form from a list of options. The difference between tests of form and meaning is that in the former, word meaning is provided in the test prompt and examinees must demonstrate knowledge of the L2 word form, whereas in the latter, the L2 form is provided, and examinees demonstrate understanding of its meaning.

Figure 1*Four Test Modalities of Form-Meaning Knowledge*

Note. The meaning-recognition item is adapted from the New General Service List Test (Stoekel et al., 2018).

Within these four modalities, several variations exist. First, as depicted in Figure 1, bilingual formats are sometimes used to reduce the risk of target word knowledge being conflated with the ability to understand other elements of the test item (e.g., Nguyen & Nation, 2011). Second, also shown in Figure 1, in form-recall tests, one or more letters of the target word are sometimes provided to limit possible correct responses to only the target item (e.g., Schmitt et al., 2021). Third, scoring of recall tests sometimes (e.g., Schmitt et al., 2021), but not always (e.g., Stoekel et al., 2019), requires correct spelling of the desired response.

A fourth difference, and a focus of the present paper, is the amount of context in the test items. Existing tests of form-meaning knowledge range from those providing the target item devoid of any context to those in which the target structure is embedded in substantial context that may aid lexi-

cal inferencing. Figure 2 depicts four levels of contextualization for each of the previously described aspects of form-meaning knowledge. These levels are admittedly somewhat arbitrary but are meant to represent important points along a continuum of possibilities. At level 1, only the target item is provided in the item stem. This level of contextualization appeared in early versions of the TOEFL (Read, 2000) and is employed in the Vocabulary Levels Test and its variants (VLT; Nation, 1983; Schmitt et al., 2001). At level 2, the stem contains a short sentence with only enough context to orient test-takers to the target item's part of speech. The Vocabulary Size Test (VST; Nation & Beglar, 2007) is an example of an instrument using this level of contextualization. Level 3 item stems are also one sentence in length while level 4 are a paragraph or more. Levels 3 and 4 differ from level 2 in that they may contain information to aid lexical inferencing. Inference-generating information may be intentionally included in all test items, as in Sasao and Webb's (2018) Guessing from Context Test. Alternatively, the presence or amount of information to aid inferencing may vary across items. Laufer's (1989) vocabulary measure is a good example of this. Her purpose was to determine which words were known in a normal reading passage, and content words were target items, whether they could be inferred from their context in the passage or not.

To our knowledge, the two highest levels of contextualization are used only in meaning-recognition tests. An example of level 3 is Sasao and Webb's (2018) aforementioned diagnostic test of lexical inferencing ability, and examples of level 4 can be found in standardized tests like the TOEFL. As displayed in Figure 2, at level 4, multiple target words can be assessed in a single prompt. This is a good way to balance the provision of extended context with practicality. The absence of more extensive contextualization in form-recognition and form-recall tests is understandable because in these modalities item stems are often in the L1, where single words can readily be understood in isolation. As for meaning-recall, there are potentially good uses for contextualized assessment at levels 3 and 4, but to date such tests are rarely, if ever, used. In the next sections we explore these possible assessment applications by taking a closer look at the two defining characteristics of the CMRT platform introduced in this paper: contextualization in vocabulary testing and meaning-recall assessment.

Figure 2
A Framework of Vocabulary Assessment Item Formats Measuring Form-Meaning Knowledge at Four Levels of Contextualization

		Levels of Contextualization			
		1	2	3	4
Meaning-Recognition	privilege	It is a <u>privilege</u> . (A) 宮殿 (B) 定型句 (C) 特権 (D) 欠点	It is a <u>privilege</u> . (A) 宮殿 (B) 定型句 (C) 特権 (D) 欠点	Having access to clean water and good food is a <u>privilege</u> that many people around the world do not have.	Having access to clean water and good food is a (1) <u>privilege</u> that many people around the world do not have. While we (2) <u>satisfy</u> our thirst with a simple turn of the (3) tap, many people face the daily challenge of finding safe water sources. It's crucial for us to recognize this (4) <u>disparity</u> and work towards ensuring everyone's right to these essential resources.
Form-Recognition	特権	それは <u>特権</u> です。 (A) palace (B) expression (C) privilege (D) defect	それは <u>特権</u> です。 (A) palace (B) expression (C) privilege (D) defect	(A) 宮殿 (B) 定型句 (C) 特権 (D) 欠点	1. (A) 宮殿 (B) 定型句 (C) 特権 (D) 欠点
Meaning-Recall	privilege = ____	It is a <u>privilege</u> . privilege = ____	It is a <u>privilege</u> . privilege = ____		
Form-Recall	特権 = pr ____	It is a pr ____ (特権)	It is a pr ____ (特権)		

Note. The example level 2 meaning-recognition item comes from the New General Service List Test (Stoeckel et al., 2018). All other example items are extrapolated from that.

Contextualization in Vocabulary Testing

Test Uses

The amount of context provided in vocabulary test items should reflect the purpose and intended consequences of testing (Read & Chapelle, 2001). Formally, Read and Chapelle distinguish between context-independent and context-dependent vocabulary assessment. In the former, the expected response can be made without reference to context while in the latter, understanding of contextual information in the test item is necessary to answer correctly. Thus, levels 1 and 2 in our framework are context-independent while levels 3 and 4 could be either, depending on whether test items can be answered without comprehending the context provided in the stem. Examples of context-dependent items might be found in a test of lexical inferencing in which the target items are pseudowords whose meanings can only be worked out from the context provided. Let us now consider the possible role of context for several vocabulary testing purposes.

Assessing Lexical Inferencing Ability

Obviously, when we wish to assess the ability to guess words from context, a context-dependent item format is indispensable. Lexical inferencing is an important vocabulary-development strategy (Nation, 2008), so there is utility in diagnostically assessing this skill and helping students become better at it (Nation, 2013).

Testing Isolated Knowledge of the Form-Meaning Link

In contrast, when we wish to measure understanding of the form-meaning link in isolation, context-independent assessment is required so that examinees are unable to employ the separate skill of guessing unknown items from context (Schmitt, 2010). Accordingly, size and levels tests such as the VST (Nation & Beglar, 2007) and VLT (Nation, 1983) are typically context-independent.

Measuring Vocabulary Knowledge for Reading

There are arguments for and against contextualized vocabulary assessment for the receptive skills. In fluent reading, context usually offers relatively little support for understanding word meaning because automaticity in word recognition and meaning retrieval is required to free up cognitive resources for text-level meaning construction (Grabe, 2009). Therefore, it is

sometimes claimed that context-independent tests are better gauges of the lexical understanding typically employed in reading (Cameron, 2002).

However, in coverage-comprehension studies, in which researchers investigate how differences in comprehension correspond with small changes in the percentage of words known in a text, there may be a case for assessing knowledge of lexis in the context of a study's reading passage. Word meaning may be understood when assessed in a non-contextualized manner but not when used with a specialized meaning in the passage (Webb, 2021b). Alternatively, a word may be understood in the supportive context of a natural text but not in a discrete point test item. Though research on previously unread text has not found a significant difference in scores on fully-contextualized (i.e., level 4) and non-contextualized vocabulary tests (Henning, 1991), just one study has examined this issue. If researchers wish to measure the precise percentage of words known in a particular text, perhaps context should be considered. Indeed, in previous coverage-comprehension research, both approaches to vocabulary measurement have been used (see Laufer, 1989 and Schmitt et al., 2011).

Testing the Assumptions of the Word Family

A word family consists of a base form (e.g., *use*) together with related inflectional (e.g., *used*, *uses*) and derivational forms (e.g., *useful*, *useless*). The precise members of a family depend on the definition used (see Bauer & Nation, 1993), but a general assumption underlying the word family is that when a learner knows the meaning of one member, they should also be able to receptively understand other members when encountered in a meaningful context (Nation, 2015). Thus, contextualized vocabulary assessment may be preferred in studies investigating this assumption of the word family (Laufer et al., 2021). Such an approach might yield different results from research that has assessed word knowledge with no supportive context and found relatively low correspondence between baseword and derivational form knowledge (e.g., Ward & Chuenjundaeng, 2009). This is uncertain, however, because, as discussed below, comparisons of tests with different levels of contextualization have yielded inconsistent findings (Henning, 1991; Laufer, 2023).

Promoting Positive Washback

Washback is the effect that tests have on teaching and learning. Although tests are probably not administered solely for their washback, selection of item format can be influenced by the perceived washback a test has (Read

& Chapelle, 2001). There are divergent views regarding the washback of context in vocabulary test items. When vocabulary items in the TOEFL were changed from discrete point to those embedded in reading passages, it was thought to bring about positive washback in that it would encourage test-takers to learn to deal with vocabulary in communicative contexts (Read & Chapelle, 2001). Similarly, Qian (2008) has stated that non-contextualized vocabulary testing can have negative washback if it encourages the study of words in isolation. Nation (2013), however, disagrees with this view, citing research that the use of word cards and other forms of limited-context study are effective for learning new words (de Groot, 2006; Elgort, 2011).

Research on Contextualization in Tests of Form-Meaning Knowledge

Several studies have directly compared levels of contextualization in assessment of form-meaning knowledge. This research has almost exclusively investigated meaning-recognition item types and paints a moderately favorable picture for the use of increased contextualization in vocabulary assessment.

Of the areas explored, two have not been impacted by differences in contextualization. The first is the correlation between vocabulary and reading test scores. This research has compared vocabulary assessment at contextualization levels 1 and 3 (Qian, 2002), 2 and 4 (Ushiro et al., 2009), and 3 and 4 (Qian, 2008). In each case, the vocabulary-reading correlation did not significantly differ for the compared vocabulary measures. Second, though only levels 3 and 4 have been compared, differences in context have not been found to influence item discrimination as estimated with point-biserial correlations (Qian, 2008). This means that test items employing the compared levels of contextualization did not differ in their capacity to distinguish learners on the basis of vocabulary knowledge.

Research has also identified two areas that have been affected by the level of contextualization in vocabulary items. The first is concurrent validity. In Henning's (1991) aforementioned comparison of meaning-recognition items at each of our four levels of contextualization, level 4 scores correlated most strongly with a criterion vocabulary measure, with the difference between levels 1 and 4 reaching significance. Second, added context may favorably impact test reliability. Henning (1991) found that estimates of internal reliability consistently increased with contextualization across tests with the same number of items. The differences were significant for level 4 in comparison to levels 1 and 2 and nearly significant relative to level 3.

Finally, research has yielded inconsistent results on the impact of changes in contextualization on item difficulty. Henning (1991) assessed the same words across five item types¹ at all four levels of contextualization and found no significant difference in mean scores. On the other hand, Laufer (2023) found a significant difference in scores when testing knowledge of the same pseudowords at three levels of contextualization. She provided learners with the meanings of 22 pseudo-basewords (e.g., *stace*) and then tested their ability to understand derivations of those words (e.g., *stacement*) at contextualization levels 1-3. Mean scores at level 1 were significantly lower than at levels 2 and 3. Perhaps these disparate findings can be explained by differences in the item stems used in the two studies. Whereas Henning's contained only the target word and (at levels 2-4) the context in which it was embedded, each of Laufer's item stems reminded test-takers to consider context, as in the following example:

If *stace* means "to participate," what does *stacement* mean in the following sentence?

Full and active *stacement* in school activities is required of all students.

Stacement means _____.

Another possible explanation is research showing that meaning-recognition formats, like those employed by Henning, mostly measure isolated vocabulary knowledge even when extensive contextualization is used (Ushiro et al., 2009).

Meaning-Recall Vocabulary Assessment

Considerations in Choosing Between Meaning-Recall and Meaning-Recognition

As with levels of contextualization, the type of form-meaning knowledge assessed ought to be guided by the purpose and intended consequences of testing (Schmitt et al., 2020). Because both meaning-recall and meaning-recognition assess receptive lexical knowledge, these two modalities are frequently compared, and decisions regarding which to use are often made by weighing practicality and accuracy. In the following paragraphs, we discuss these two factors together with a third consideration, washback.

Practicality refers to the ease with which tests are designed, administered, and scored (Brown, 2004). Regarding design, meaning-recall tests are clearly more practical owing to the time and expertise needed to write

good distractors for meaning-recognition tests (Rodriguez, 2005). For test administration, however, meaning-recognition is quicker (McLean et al., 2020) because test-takers only select responses rather than translate target-words. Regarding scoring, meaning-recognition is also quicker – indeed, it is instantaneous in computer-administered tests. As previously mentioned, the scoring of meaning-recall tests has become easier with automated marking, but currently novel responses still require human attention. When there are numerous examinees or when results are needed quickly, meaning-recognition tests remain the more practical option. However, for classroom assessment purposes such as achievement tests, and for many research applications, any difference in test practicality may be outweighed by considerations of accuracy and washback.

The accuracy of a language test is based on its capacity to (a) detect knowledge when it is present and (b) detect the absence of knowledge when it is absent. These are referred to as sensitivity and specificity, respectively (Eckes, 2017). It is sometimes claimed that meaning-recognition is more sensitive than meaning-recall (Webb, 2021a), as evidenced by the many studies showing that learners achieve higher scores on meaning-recognition tests (e.g., Kremmel & Schmitt, 2016; Laufer & Goldstein, 2004; Stoeckel et al., 2019; Stoeckel & Sukigara, 2018). However, meaning-recognition tests are influenced by the use of construct-irrelevant test strategies and blind guessing (Gyllstad et al., 2015; McDonald, 2015; McLean, Kramer, & Stewart, 2015), indicating that a portion of the score difference between the two test formats is due to decreased specificity rather than increased sensitivity of the meaning-recognition measure. Hence, some scholars consider meaning-recall to be the more accurate of the two test formats, at least as a measure of the lexical knowledge used in reading (Kremmel & Schmitt, 2016; McLean, 2021; Schmitt, 2019; Stoeckel et al., 2021). Perhaps an indication of how widely this second view is held, meaning-recall tests are commonly employed as criterion measures in validation studies of meaning-recognition tests (e.g., Kremmel & Schmitt, 2016; Stoeckel et al., 2019; Webb et al., 2017), but rarely, if ever, the other way around.

An overlooked factor favoring the use of meaning-recall is washback. Compared to meaning-recognition, meaning-recall is a stronger form of lexical knowledge (Laufer & Goldstein, 2004) that correlates more strongly with receptive language ability (McLean et al., 2020; Zhang & Zhang, 2022). There is, therefore, good reason to encourage learners to master vocabulary to the level of meaning-recall, and perhaps meaning-recall vocabulary assessment would have that effect. While this may be difficult to enact in large-

scale educational testing, it should be considered for smaller-scale uses like classroom progress tests and quizzes.

Research Comparing Meaning-Recall and Meaning-Recognition

The studies comparing meaning-recall and meaning-recognition vocabulary measures have produced relatively consistent findings. First, meaning-recall has better internal reliability. Though statistical significance has gone unreported, this has been found in each study that reported the reliability of both measures and that assessed the same words under the two item formats (McLean et al., 2020; Stoeckel et al., 2019; Stoeckel & Sukigara, 2018). Second, meaning-recall appears to be a better predictor of reading comprehension. Although some studies have lacked statistical significance (Jeon & Yamashita, 2014; Laufer & Aviad-Levitzky, 2007), others have found a clear contrast (Zhang & Zhang, 2022) with large effect sizes (McLean, et al., 2020). Third, as previously stated, meaning-recall tests require more time to administer. Note, however, that this difference has reached statistical significance for multiple-choice but not matching formats (McLean et al., 2020). Fourth, when the same words are tested, meaning-recall is more difficult than meaning-recognition (Kremmel & Schmitt, 2016; Laufer & Goldstein, 2004; Stoeckel et al., 2019; Stoeckel & Sukigara, 2018). Related research has indicated that reasons for this difference include random guessing and use of not only construct-relevant but also construct-irrelevant test strategies on the meaning-recognition test (Gyllstad et al., 2015; McDonald, 2015; McLean, Kramer, & Stewart, 2015).

In sum, while meaning-recall tests may be somewhat less practical, they are more accurate, a better predictor of receptive language ability, and – we would argue – more likely to produce beneficial washback. In the final section, we provide a detailed description of meaning-recall assessment on the CMRT platform.

The Contextualized Meaning-Recall Testing Platform

The CMRT platform (<https://cmrt.vocableveltest.org>) can be used by L2 teachers and researchers to expeditiously create, administer, and mark contextualized meaning-recall vocabulary tests. For test creation, the platform is set up so that anyone with a Gmail account can create and administer tests. The test owner simply inputs a text and selects target words or phrases. This produces a test in which learners see discourse with boxes under the target items to input their responses. Although this format enables examinees to

Figure 3
Example Test Items at Four Levels of Contextualization on the CMRT Platform

Level 1

privilege

satisfy

tap

disparity

Level 2

It is a privilege .

Level 3

Having access to clean water and good food is
 a privilege that many people around the world

 do not have.

Level 4

Having access to clean water and good food is
 a privilege that many people around the world

 do not have. While we satisfy our thirst

 with a simple turn of the tap , many

 people face the daily challenge of finding
 safe water sources. It's crucial for us to
 recognize this disparity and work towards

 ensuring everyone's right to these essential
 resources.

consider broad context when discerning word meaning, the platform can be used to assess vocabulary at all four levels of contextualization, as shown in Figure 3. There is also a place for test creators to input instructions. This allows for the elicitation of different kinds of responses (e.g., L1 translations; L2 synonyms, definitions, explanations) depending on the learner group and testing purpose. To administer a test, the test creator needs only to share the test URL with test-takers. Examinees do not need to register as members of the site. To deter students from getting outside help, there is also an option to first warn test-takers and then automatically end the test if navigation away from the test app is detected. Regarding privacy, the platform is hosted on a secure cloud server, and if an added layer of protection is desired, students could be asked to use pseudonyms or examinee codes instead of their actual names.

After test administration, either the owner or one or more assigned raters mark the test. Raters access a list of distinct responses for each item and rate them (as correct, incorrect, or partially correct) without seeing the judgments of other raters (Figure 4). When marking is complete, all judgments can be viewed and final decisions recorded for discrepant ratings. Additionally, these decisions can be saved for future test use, reducing the burden of marking in subsequent test administrations. The test owner can also view and download tables of responses and points earned to each test item for every examinee (Figures 5 and 6).

Figure 4
Rater Interface for the CMRT Platform

Item 1: kind
...Have you heard of aerobics? Aerobics is a KIND of exercise in which you move your body a lot so ...

● やさしい

Incorrect Partially Correct Correct

Comments

● 一種の

Incorrect Partially Correct Correct

Comments

● 種類

Incorrect Partially Correct Correct

Comments


Note. Each target word is displayed followed by (a) the context in which it appears in the passage (in small grey font) and (b) a list of each distinct response (in pink boxes), which can be rated as incorrect, partially correct, or correct. There is also a place for comments for ambiguous or otherwise noteworthy responses.

Figure 5
Online View of Test Responses

Responses													Answers	Scores				
Name	Start Time	Finish Time	Total Time	kind	heart	healthy	aerobics	hard	routine	powerful	judge	Participants	backgrounds	pair	spread	championship	International Olympic Committee	official
Mike M.	2022-08-20 10:30:26	2022-08-20 10:36:20	00:05:54	やさしい	心	健康	エアロビクス	激しく	ルーティーン	パワフル	審判	参加者	バックグラウンド	ペア	広がった	チャンピオン	国際オリンピック委員会	公式の
ky	2022-08-20 10:39:01	2022-08-20 10:43:05	00:04:04	種類	心臓	健康	エアロビクス	一生懸命	ルーティーン	強い	審判	参加者	音楽、経験	ペア、一足	広がった	大会	国際オリンピック委員会	公式の
Kodani	2022-08-20 12:24:29	2022-08-20 12:33:28	00:08:59	種類の	心臓	健康に	エアロビクス	がんばる	ノルマ	力強く	裁判官	関心を持つ人	豊かな経験	*訳さな	ひろがった	選手権	国際オリンピック委員会	公式の
Urara	2022-08-20 14:15:19	2022-08-20 14:20:07	00:04:48	種類	心臓	健康	エアロビクス	がんばる	ルーティーン	力強い	審査員	競技者	経験	一足	広がる	大会	国際オリンピック委員会	正式
tomoko	2022-08-20 15:45:53	2022-08-20 15:48:56	00:03:03	種類	心臓	健康な	エアロビクス	熱心に	ルーティーン	力強い	審判	参加者	背景	ひと揃い	広まる	競技会	国際オリンピック委員会	公式な

Note. Responses can be viewed online and downloaded as a csv file.

Figure 6
 Online View of Ratings for Test Responses

Responses Answers Scores 

Name	Start Time	Finish Time	Total Time	kind	heart	healthy	aerobics	hard	routine	powerful	judge	Participants	backgrounds	pair	spread	championship	International Olympic Committee	official
Mike M.	2022-08-20 10:30:26	2022-08-20 10:36:20	00:05:54	0	0.5	1	1	1	1	1	1	1	0.5				1	
ky	2022-08-20 10:39:01	2022-08-20 10:43:05	00:04:04	1	1	0.5	1	1	1	1	1	1	1	1			1	
Kodani	2022-08-20 12:24:29	2022-08-20 12:33:28	00:08:59	1	1	1	1	1	0	1	1	0	1				1	
Urara	2022-08-14 15:19	2022-08-20 14:20:07	00:04:48	1	1	0.5	1	1	1	1	1	1	1					
tomoko	2022-08-15 45:53	2022-08-20 15:48:56	00:03:03	1	1	1	1	1	1	1	1	1	0.5				1	

Note. Blank cells depict unrated items; (0 = incorrect, 0.5 = partially correct, 1 = correct). Ratings can be viewed online or downloaded as a csv file.

Although the CMRT platform can substantially reduce the amount of time needed for meaning-recall assessment at several levels of contextualization, it has limitations. As a practical matter, because the platform was developed with grant funding for a particular research project, users should expect limited technical support. Also, tests can be rendered only in left-justified, plain text. Formatting options like bold or italicized font, underlining, centering, and auto-numbering are unavailable. Moreover, although test results are downloadable as csv files, the platform is not integrated into any existing learning management system. Concerning construct validity, it should not be assumed that tests developed and administered on the CMRT platform are valid for particular purposes. For low-stakes, classroom use, teachers can probably apply the same principles they use for other forms of assessment. For higher-stakes testing or research, however, validation evidence must be gathered to support the use of CMRTs for specific uses and score interpretations (Messick, 1995). Moreover, when multiple target words occur in close proximity in a single passage, the assumption of local independence would need to be checked (de Ayala, 2009).

Conclusions and Future Directions

The CMRT platform has potential uses in both research and pedagogy. In research, it could be employed to compare meaning-recall vocabulary assessment at different levels of contextualization, paralleling the above-mentioned studies on meaning-recognition formats. Additionally, inquiry comparing meaning-recognition and meaning-recall might be extended to more systematic investigation of the role of context. Coverage-comprehension studies could also be conducted with vocabulary measured at the meaning-recall level (see McLean, 2021). For teachers, the CMRT platform may be used for practicing and assessing lexical inferencing ability, where it is advantageous to provide learners with a continuous text with actual target words rather than blanks or pseudowords (Nation, 2013). Additionally, since learners tend to achieve higher scores when vocabulary is tested in the same context in which it is learned (Watanabe, 1997), the platform could be used as a sensitive measure of newly acquired vocabulary from class texts. Finally, its use in the classroom may promote positive washback if it encourages students to study and learn words more deeply than they would with meaning-recognition achievement tests.

In closing, we would like to address a reviewer's intriguing question regarding the possible use of AI to judge test responses. With the recent, rapid development of generative AI capabilities, it would be interesting to see

whether this is feasible. A foreseeable challenge is that there are multiple ways to express word meaning that go well beyond dictionary definitions and one-to-one translations. Humans can achieve high levels of inter-rater reliability for such responses, even when using rather nuanced marking criteria. Whether AI could match humans in this regard is an interesting question for researchers to explore.

Notes

1. Henning (1991) compared eight item types in total, but only five strictly assessed form-meaning knowledge. Of these, there was one item at each of contextualization levels 1, 2, and 4, and two items at level 3.

Acknowledgements

This study was partially funded by grant number JP 22K00793 from the Japan Society for the Promotion of Science.

Tim Stoeckel teaches English and applied linguistics at the University of Niigata Prefecture. His primary interests are in L2 reading, vocabulary learning and assessment, and the relationship between lexical knowledge and reading comprehension.

Stuart McLean holds doctorates in medicine and TESOL, and a PGCE. He publishes research related to vocabulary, assessment, and comprehension (https://scholar.google.com/citations?hl=en&user=yL_1NxsAAAAJ). His on-line self-marking meaning-recall (reading or listening) levels tests are at www.vocableveltest.org. Teachers can download automatically marked responses, typed responses, and the time taken to complete responses.

Paul Raine (MA TEFL/TESL) is an award-winning teacher, presenter, author, and developer. He has published numerous research articles on the teaching and learning of English as a second language, and is particularly interested in Computer Assisted Language Learning (CALL). He currently lives and teaches English in Osaka, Japan.

Hung Tan Ha is a PhD student in Applied Linguistics at Victoria University of Wellington. His research interests involve L2 reading comprehension and vocabulary assessment.

Nam Thi Phuong Ho teaches English Language at the School of Foreign Languages - University of Economics Ho Chi Minh City. Her research inter-

est is English for Specific Purposes, TESOL Methodologies, Quantitative and Qualitative research designs, and Vocabulary teaching and learning.

Young Ae Kim is an instructor at Kindai University and Kyoto Seika University and holds an MA from Kansai University. She is working on research related to test item format type, word counting units, listening and word difficulty, as well as making the items for the tests available at www.vocableveltest.org.

References

- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279. <https://doi.org/10.1093/ijl/6.4.253>
- Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Longman.
- Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, 6(2), 145–173. <https://doi.org/10.1191/1362168802lr103oa>
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.
- de Groot, A. M. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, 56(3), 463–506. <https://doi.org/10.1111/j.1467-9922.2006.00374.x>
- Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing*, 34(3), 383–411. <https://doi.org/10.1177/0265532216672703>
- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61(2), 367–413. <https://doi.org/10.1111/j.1467-9922.2010.00613.x>
- Grabe, W. (2009). *Reading in a second language*. Cambridge University Press.
- 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, 278–306. <https://doi.org/10.1075/itl.166.2.04gyl>
- Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items*. Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1991.tb01390.x>

- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64, 160–212. <https://doi.org/10.1111/lang.12034>
- Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24(4), 617–637. <https://doi.org/10.1017/S0272263102004047>
- 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>
- Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.
- Laufer, B. (2023). Understanding L2-derived words in context: Is complete receptive morphological knowledge necessary? *Studies in Second Language Acquisition*, 1-14. <https://doi.org/10.1017/S0272263123000219>
- Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning-recall or word meaning-recognition? *The Modern Language Journal*, 101, 729–741. <https://doi.org/10.1111/modl.12431>
- Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436. <https://doi.org/10.1111/j.0023-8333.2004.00260.x>
- Laufer, B., Webb, S., Kim, S. K., & Yohanan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL-International Journal of Applied Linguistics*, 172(2), 229–258. <https://doi.org/10.1075/itl.20020.lau>
- McDonald, K. (2015). The potential impact of guessing on monolingual and bilingual versions of the vocabulary size test. *Osaka JALT Journal*, 2, 44–61. <https://bit.ly/3PoRniq>
- McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language*, 33, 126–140. <https://nflrc.hawaii.edu/rfl/item/528>
- McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, 19(6), 741–760. <https://doi.org/10.1177/1362168814567889>

- McLean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, 4, 26–35. <https://doi.org/10.7820/vli.v04.1.mclean.et.al>
- McLean, S., Raine, P., Pinchbeck, G., Huston, L., Kim, Y. A., Nishiyama, S., & Ueno, S. (2021). The internal consistency and accuracy of automatically scored written receptive meaning-recall data: A preliminary study. *Vocabulary Learning and Instruction*, 10(2), 64–81. <https://doi.org/10.7820/vli.v10.2.mclean>
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 37, 389–411. <https://doi.org/10.1177/0265532219898380>
- 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>
- Nation, I. S. P. (1983). Teaching and testing vocabulary. *Guidelines*, 5, 12-25.
- Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Heinle ELT.
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139858656>
- Nation, P. (2015). Which words do you need? In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 568–581). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199641604.013.016>
- Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31, 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7
- Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual Vocabulary Size Test of English for Vietnamese learners. *RELC Journal*, 42(1), 86–99. <https://doi.org/10.1177/0033688210390264>
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–536. <https://doi.org/10.1111/1467-9922.00193>
- Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly*, 5(1), 1–19. <https://doi.org/10.1080/15434300701776138>
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732942>
- Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1–32. <https://doi.org/10.1177/026553220101800101>

- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3–13. <https://doi.org/10.1111/j.1745-3992.2005.00006.x>
- Sasao, Y., & Webb, S. (2018). The guessing from context test. *ITL-International Journal of Applied Linguistics*, 169(1), 115–141. <https://doi.org/10.1075/itl.00009.sas>
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. <https://doi.org/10.1057/9780230293977>
- Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, 52(2), 261–274. <https://doi.org/10.1017/S0261444819000053>
- Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, 12(4), e622. <https://doi.org/10.1002/tesj.622>
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- 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>
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55–88. <https://doi.org/10.1177/026553220101800103>
- Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2023). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*. Advance online publication. <https://doi.org/10.1177/02655322231162853>
- Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. G. (2021). What the research shows about written receptive vocabulary testing: A reply to Webb. *Studies in Second Language Acquisition*, 43(2), 462–471. <https://doi.org/10.1017/S0272263121000437>
- Stoeckel, T., Ishii, T., & Bennett, P. (2018). A Japanese-English bilingual version of the New General Service List Test. *JALT Journal*, 40(1), 5–21. <https://doi.org/10.37546/JALTJ40.1-1>

- Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 43(1), 181–203. <https://doi.org/10.1017/S027226312000025X>
- Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning-recall vocabulary knowledge. *System*, 87, 102161. <https://doi.org/10.1016/j.system.2019.102161>
- Stoeckel, T., & Sukigara, T. (2018). A serial multiple-choice format designed to reduce overestimation of meaning-recall knowledge on the Vocabulary Size Test. *TESOL Quarterly*, 52, 1050–1062. <https://doi.org/10.1002/tesq.429>
- Ushiro, Y., Hirai, N., Hoshino, Y., Nakagawa, C., Kai, A., Ide, M., Oki, T., Suzuki, S., Terada, Y., Takaki, S., & Shimizu, H. (2009). Comparing effects of two types of vocabulary knowledge on six question types in reading tests among Japanese EFL learners. *JLTA Journal Kiyō*, 12, 104–115. https://doi.org/10.20622/jltaj.12.0_104
- Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, 37, 461–469. <https://doi.org/10.1016/j.system.2009.01.004>
- Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, 19, 287–307. <https://doi.org/10.1017/S027226319700301X>
- Webb, S. (2021a). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 43(2), 454–461. <https://doi.org/10.1017/S0272263121000449>
- Webb, S. (2021b). Lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33(2), 278–293. <http://hdl.handle.net/10125/67407>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, 168(1), 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4), 696–725. <https://doi.org/10.1177/1362168820913998>