OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 13, Number 1 (2024)
https://doi.org/10.29140/vli.v13n1.1233

Castledown

An Exploratory Criterion Validation of Three Meaning-Recall Vocabulary Test Item Formats

Tim Stoeckel
University of Niigata Prefecture, Japan
Department of International Studies and Regional Development
stoeckel@unii.ac.jp

Tomoko Ishii
Meiji Gakuin University, Japan
Center for Liberal Arts
ishii@gen.meijigakuin.ac.jp

Abstract

In an upcoming coverage-comprehension study, we plan to assess learners' meaning-recall knowledge of words as they occur in the study's reading passage. As several meaning-recall test formats exist, the purpose of this small-scale study ($N = 10$) was to determine which of three formats was most similar to a criterion interview regarding mean score and the consistency of correct/incorrect classifications (match rate, $k = 30$). In Test 1, the prompt consisted of only the target item, and a written translation of its meaning was elicited. In Test 2, the prompt was a short sentence in which a target item was highlighted, and a written translation of only that target item was requested. In Test 3, the prompt was the same sentence as in Test 2, but the target item was unhighlighted, and participants were requested to translate the entire sentence. Finally, in the criterion interview, participants were asked to demonstrate their understanding of the target items in the same prompt sentences as in Tests 2–3. The results indicated that Test 3 produced a mean score and match rate most similar to the interview, followed by Test 2, with Test 1 being the least similar. The paper discusses several factors explaining differences in test performance that were explored during the interview.

Keywords: vocabulary assessment, meaning-recall, criterion validation.

Data Availability Statement: All relevant data are within this paper.

There are various forms of vocabulary assessment, and choices must be made depending on the purpose and intended consequences of testing (Schmitt et al., 2020). In an upcoming coverage-comprehension study, we plan to test learners' meaning-recall knowledge of the meaning of words as they occur in the study's reading passage. As some target words do not take on common dictionary meanings, we were unsure whether a standard approach to meaning-recall assessment, in which learners are provided with a sentence-length prompt and asked to translate only the highlighted target word, would be the most suitable for our purposes. Consider the following prompt:

burst: She **burst** into song.

If students provide a translation for *burst* only, a response indicating the core sense of the word is ambiguous, not necessarily indicating an ability to understand the metaphoric sense. To address such possibilities, we could have participants translate the entire sentence, with the target item un-highlighted:

She burst into song.

There are several advantages to this approach. First, participants are probably more likely to consider the meaning of the entire sentence when formulating their response, which may help them tune into the assessed meaning of the target word. It also allows for a wider range of natural L1 responses, especially when target items can be construed as something other than a one-to-one translation. For example, *afford* does not directly translate to a single Japanese word, requiring something like, "I don't have enough money to…", which might pose difficulty in a word-only translation task. Another possible advantage is that for phrasal expressions (e.g., *put down*), there is no highlighting to inform test-takers that the target phrase is a single entity. Additionally, full-sentence translations clarify for markers whether responses demonstrate an understanding of target items as used in particular contexts. Finally, there is research, albeit implied from indirect comparisons, showing that learners achieve higher scores on a full-sentence translation task than on a single-word translation task, suggesting that the former enables learners to better demonstrate their meaning-recall knowledge (Stoeckel, Ishii, et al., 2023).

However, full-sentence translations require more time for both test-takers and markers compared to single-word translations. If the difference between the two formats is minor, the more practical one may be preferable. This issue of test practicality led us to also consider a third format in which the prompt consists of only the target word. Obviously, when a test-taker translates the word *burst* from a single-word prompt, we should not expect them to provide a metaphoric sense of the word. However, if the overwhelming majority of learners who demonstrate knowledge of the core meaning-sense in response to a single-word prompt also understand the metaphorical meaning of the word in the phrase *burst into song*, time and effort could be saved by using the simpler test format.

In addition to this question of item format, we were unsure of how strictly to mark responses. Many studies that employ meaning-recall vocabulary tests use strict, dichotomous scoring. However, Kremmel and Schmitt (2016, p. 386) found that such

an approach resulted in cases where written meaning-recall responses were marked as incorrect even though learners could demonstrate knowledge of the same words in an interview. This suggests that strict scoring of written meaning-recall may underestimate word knowledge. An alternative approach may be to mark somewhat leniently, assuming that demonstrations of partial knowledge would be indications of fuller understanding.

With these considerations in mind, the purpose of the present study was to compare the three previously described written meaning-recall test formats — marked with strict, lenient, and sensitive scoring — to a criterion interview measure. The research questions (RQs) were:

1. Which combination of written test format and marking criteria yields scores most similar to those on the interview?
2. Which combination of written test format and marking criteria has the highest rate of correspondence in correct/incorrect classifications with the interview?
3. Do any word or test characteristics help explain the answers to questions 1–2?

Methods

Instruments

Target words were selected based on the results of a yes/no test of 101 words and phrases administered to 44 learners like those in the present study. Thirty items were chosen to represent a range of non-extreme item facilities. Knowledge of these target words was measured using four tests as follows (full tests are in Supplementary Materials, https://osf.io/cz73u/?view_only=16b19b5fc5d845cebca7c92c7609cb4d).

Test 1 – Word translation in response to single word stimuli

The prompt consisted of only the target item. Participants provided a written translation of its meaning.

Test 2 – Word translation in response to sentence stimuli

The prompt consisted of a short sentence containing the highlighted target item. Participants provided a written translation of only the target item while carefully considering the sentence context.

Test 3 – Sentence translation in response to sentence stimuli

The prompt consisted of the same sentence as in Test 2, but the target item was unhighlighted. Participants provided a written translation of the entire sentence.

Test 4 – Interview

In the interview, participants were asked whether they knew the meaning of the word and then to translate the same prompt sentence as in Tests 2-3. This led to follow-up questions when it was unclear whether the assessed meaning of the word was known.

Participants

The participants were 10 students in their second to fourth year at a university in the Tokyo area and were recruited, with a small monetary incentive, from a course taught by one of the researchers. Their ages were 19–22, with a variety of majors including economics, literature, and international studies.

Procedures

The participants took Tests 1–3, in order, on the Contextualized Meaning-Recall Test platform (https://cmrt.vocableveltest.org; Stoeckel, McLean, et al., 2023). This sequence was thought to minimize any testing effect. Test 1 was first because context was unavailable in the prompts. Test 3 was last because it was assumed that learners would engage most deeply in its full-sentence translation task, which could influence subsequent test-taking. Test logs confirmed that no one switched between browser tabs, and continuous monitoring via Zoom video chat gave no indication that participants utilized other resources during the tests. After Tests 1–3, there was a Zoom interview where knowledge of the target words and comprehension of the prompt sentences in Tests 2–3 were checked. When responses did not clearly demonstrate understanding of target words, follow-up questions were asked to determine precisely what was known. During this process, the interviewer had no access to the responses from Tests 1–3, though some participants voluntarily shared how they had responded to those tests. When participants did not know word meaning, the interviewer provided a quick explanation, which sometimes derived further insights from the participants.

The interviewer then marked the anonymized responses to Tests 1–3. For Test 1, any dictionary meaning of the tested word was considered correct. For Tests 2–3 (like the interview), a response was marked as correct only if it matched a plausible meaning for the prompt sentence. Accordingly, the marking criteria for Test 1 sometimes differed from that of the other tests. A question addressed by this study, then, was whether Test 1 would perform similarly to the interview despite the different marking criteria. Responses were initially marked with sensitive scoring (correct = 1, partially correct = 0.5, incorrect = 0). Partially correct marks were assigned to responses indicating only a vague understanding and those given in a wrong part of speech; such responses were then recoded as correct and then again as incorrect, to produce two additional datasets with lenient and strict scoring, respectively (for details, see Supplementary Materials, https://osf.io/cz73u/?view_only=16b19b5fc5d845cebca7c92c7609cb4d).

Inter-rater reliability was not established with the current dataset, which is a limitation of this study. However, the rater, an applied linguist and highly proficient L2 user of English, achieved a high degree of agreement with a second rater, who has a similar background to the first, on a dataset derived from a similar group of 51 learners and 14 of the same items (agreement rate = .938, Cohen's Kappa = .865). This provides some evidence of the consistency of judgments in the present study.

For RQ1, only descriptive statistics are reported because the study lacked sufficient power for meaningful tests of difference. For RQ2, confidence intervals (CIs) for the difference between proportions for paired samples (Altman et al., 2013, p. 52) were used to test for differences in the correspondence rate of correct/incorrect

classifications between each written test and the interview. Applying a Bonferroni correction, significance was set at .005, and 99.5% CIs were calculated. CIs crossing zero were considered nonsignificant.

Results and Discussion

Table 1 provides descriptive statistics for the four measures. In this small dataset, there were responses marked as partially correct only in Tests 1–2. Accordingly, in Tests 3–4, the different scoring criteria produced the same results. Regarding RQ1, mean scores gradually increased across the four tests. In Tests 1–2, lenient scoring produced scores most similar to the interview, but the difference was very small for Test 2. Among the three written tests, Test 3 (M = 21.1) scores were closest to those of the interview (M = 21.4).

Addressing RQ2, Table 2 shows that among the written tasks, Test 3 had the highest rate of correspondence of correct/incorrect classifications with the interview. (Sensitive scoring is not displayed because, with partial-credit categories, it is not directly comparable to the dichotomously scored strict and lenient datasets.) As the bottom of Table 2 indicates, the match rates for Test 1 were significantly lower than those for Tests 2 and 3, regardless of the marking criteria. The match rates for Tests 2 and 3 did not differ significantly (see Supplementary Materials for additional data required for these tests of significance.)

To further clarify RQs 1 and 2, Table 3 provides item facility values and match rate statistics for individual target words under lenient scoring (the approach most similar to the interview). The bottom of the table shows these statistics for multi-word expressions, metaphorically-used words, and all remaining target items separately. As discussed below, this empirical data, together with the interview, provided some clues as to why certain written formats differed from the interview (RQ3).

Multi-word expressions

Some items were composed of multiple words (e.g., *as if*, *put down*), which may have caused some confusion in Test 1. In the case of *as if*, some participants responded *moshi* (*if*), ignoring the word *as*, even though they later had no problem comprehending

Table 1 *Descriptive Statistics*

	1. Word			2. Word in Sentence			3. Sentence	4. Interview
	Strict	Sensitive	Lenient	Strict	Sensitive	Lenient		
Min	9	9.5	10	13	14	15	17	18
Max	24	24	24	27	27	27	27	28
M	18.30	18.95	19.60	19.90	20.00	20.10	21.10	21.40
SD	4.86	4.70	4.60	4.23	4.06	3.90	3.25	3.13
Alpha	0.82	0.84	0.83	0.81	0.81	0.78	0.67	0.70

Table 2 *Matrices for Match Rates with the Interview*

Outcome	1. Word		2. Word in sentence		3. Sentence
	Strict	Lenient	Strict	Lenient	
Response Pattern					
Both Correct	158	168	195	196	205
Interview Correct	56	46	19	18	9
Written Test Correct	25	28	4	5	6
Both Incorrect	61	58	82	81	80
Match Rate	0.730	0.753	0.923	0.923	0.950
99.5% CI for Difference Between Match Rates					
1. Word (strict)		[–0.01, 0.06]	**[0.12, 0.27]**	**[0.12, 0.27]**	**[0.14, 0.29]**
1. Word (lenient)			**[0.10, 0.24]**	**[0.10, 0.24]**	**[0.12, 0.27]**
2. Word in Sentence (strict)				[–0.03, 0.03]	[–0.02, 0.07]
2. Word in Sentence (lenient)					[–0.02, 0.07]

Note. Bold indicates statistical significance.

the meaning of this phrase in a sentence. This seemed to be because of how the item was presented rather than insufficient knowledge. On the other hand, some participants thought that the multi-word expression *put down* could not simply mean "put something down" because they knew other idioms with a variety of meanings (e.g., *put off*, *put up with*). This led to some creative (but incorrect) ideas as to what *put down* might mean, lowering Test 1 scores. However, when they encountered the same expression in a sentence in Test 2 (*Please put that down*), some participants felt comfortable responding in a straightforward manner. Some, though, remained hesitant and were surprised to learn the correct answer in the interview.

Metaphorically used words

Metaphorically used words showed the opposite trend to multi-word units, with the largest number of participants responding correctly in Test 1. This is because participants could receive a correct mark for providing a core meaning for such items seen in isolation in Test 1, but they could not extend this to demonstrate understanding of the words used metaphorically in Tests 2–4.

Katakana

Katakana is a component of the Japanese writing system used, among others, to render loanwords into the Japanese phonological system (e.g., *operator* as *opereetaa*; Daulton, 2008). We intentionally avoided mentioning katakana in the test instructions; explicitly allowing it may encourage the phonological guessing of unknown words

Table 3 *Item Statistics*

Target item	Item facility				Match rate with the interview		
	Word	Word in sentence	Sentence	Interview	Word	Word in sentence	Sentence
as if*	.2	.7	.8	.8	.4	.9	1.0
burst**	.7	.2	.3	.3	.6	.9	1.0
circle	.9	1.0	1.0	.9	.8	.9	.9
cover	1.0	.9	1.0	1.0	1.0	.9	1.0
crash	1.0	.9	1.0	.9	.9	1.0	.9
dead	.9	1.0	1.0	1.0	.9	1.0	1.0
escaped	.9	.8	.8	.8	.9	1.0	1.0
even	.6	.6	.6	.6	.8	1.0	1.0
find**	1.0	.7	.8	.8	.8	.9	1.0
firm	.1	.1	.0	.1	1.0	.8	.9
fright	.0	.1	.1	.1	.9	1.0	1.0
getting	1.0	1.0	1.0	1.0	1.0	1.0	1.0
just	.9	.8	.9	.8	.7	.8	.9
like	1.0	1.0	1.0	.9	.9	.9	.9
make	1.0	1.0	1.0	1.0	1.0	1.0	1.0
matter	.8	.9	1.0	1.0	.8	.9	1.0
meanwhile	.6	.6	.6	.6	1.0	1.0	1.0
operator	.5	.3	.5	.8	.5	.5	.7
patient	.9	.7	.7	.7	.8	1.0	1.0
pause	.4	.6	.6	.6	.8	1.0	1.0
pick up*	.3	.8	.8	1.0	.3	.8	.8
put down*	.5	.7	.7	.8	.7	.9	.9
rang	.3	.8	.8	.9	.4	.9	.9
receiver	.7	.3	.4	.3	.6	1.0	.9
reply	.6	.9	.9	.9	.7	1.0	1.0
round	.5	.0	.1	.1	.4	.9	1.0
struck**	.4	.2	.2	.2	.8	1.0	1.0
suspicion	.1	.5	.6	.6	.5	.9	.8
telephoned	1.0	1.0	1.0	1.0	1.0	1.0	1.0
thought	.8	1.0	.9	.9	.7	.9	1.0
multi-word expressions (k = 3)	.333	.733	.767	.867	.467	.867	.900
metaphorically used words (k = 3)	.700	.367	.433	.433	.733	.933	1.000
all others (k = 24)	.688	.700	.729	.729	.792	.929	.950
ALL (k = 30)	.653	.670	.703	.713	.753	.923	.950

Note. * = multi-word expression; ** = item used metaphorically in tests 2–4.

while prohibiting it may impede the easiest way to demonstrate understanding. The interview revealed how some students were reluctant to use katakana in the written tests. They explained that providing a phonological representation seemed like cheating and reported not getting credit for katakana test responses in secondary school. In the case of the word *operator*, only one participant answered with katakana and just two others provided acceptable Japanese translations, yet nine recognized the word in the interview, and eight demonstrated a good comprehension of the sentence. Participants seemed more willing to provide katakana responses in the interview because the interviewer reminded them to demonstrate even partial knowledge, and because participants had an opportunity to give further explanations after providing an initial response which, if given in katakana, may have felt like cheating. However, for *receiver*, of the three learners who used katakana, just two demonstrated understanding of the word in the interview. These observations suggest that katakana may be a source of some imprecision in written meaning-recall tests with Japanese learners.

Word misrecognition with no-context prompts

Presenting the items in isolation caused some word misrecognition. For instance, although *rang* did not pose much problem when embedded in a sentence (*It rang three times*), many participants mistook it for *lung* in Test 1. The same issue appeared with *fright* (confused with *flight*) and *thought* (*though*). Seeing words in a sentence seemed to have helped participants recognize the words more accurately, perhaps due to the provision of additional meaning and clues as to the target word part of speech. Consequently, Test 1 yielded the lowest scores even though a wider range of responses was possible for some items.

Meaning-restricting context

In general, embedding target words in limited-context sentences supported meaning recall. For instance, *suspicion* had a higher item facility when tested in a sentence, with participants explaining how the context (*She looked at me with suspicion*) was helpful in recalling the meaning of the previously-learned word. However, learners sometimes struggled when the sentence context indicated an unfamiliar meaning for an otherwise known word. For instance, some participants who correctly demonstrated knowledge of *burst* as *bakuhatsu suru* (to explode) in Test 1 could understand neither the use of this word nor the prompt sentence as a whole in *She burst into song* in Tests 2–4.

Accuracy & internal reliability

There was an interesting interplay between accuracy and internal reliability. If we accept Schmitt's (2010, p. 182) notion that a spoken interview is probably the most accurate way to discern whether learners know word meaning, then the increasing scores across Tests 1–4 together with the increasing match-rate with the interview across Tests 1–3 indicate that the order of tests by accuracy was *opposite* the order of tests by internal reliability (Table 1). Although the small samples might limit the extent to which we can argue this with confidence, this is a reminder that internal reliability coefficients should be interpreted with all relevant facts in mind. Alpha is

likely to have decreased across the four tests because as scores increased, the range of scores and standard deviations decreased, making it increasingly difficult to reliably separate learners by scores. As such, the lower alpha coefficients would not indicate poorer test quality. Regarding accuracy, another important observation is that learners sometimes demonstrated knowledge of a word in a written test but failed to do so in the interview. Such cases are expected for Test 1, which had different marking criteria, but they also occurred in Tests 2–3 (shown in the "Written Test Correct" row of Table 2). This shows that although an interview produces the highest scores and is therefore the method in which learners are most likely to demonstrate actual knowledge, it is not infallible, as neither test-takers nor interviewers perform exactly the same with each test administration.

Conclusions and Future Directions

We set out to determine which of three written meaning-recall item formats would be most suitable for coverage-comprehension research. Because we wish to discern whether learners understand specific words as they appear in a particular reading passage, Test 1 seems inappropriate. It yielded scores 6% lower than those in the interview and had a match rate with the interview of just 75.3%. Tests 2 and 3 are both strong candidates in that they produced similar scores and similarly high match rates with the interview. However, considering that even small differences in estimated coverage are consequential in coverage-comprehension research, the 3% difference in raw scores and 2.7% difference in match-rates provide support for using a sentence translation task, which, for both metrics, was more similar to the criterion measure. The findings also support the use of lenient marking (i.e., giving full credit for demonstrations of partial knowledge), as its use yielded results most similar to those of the interview.

This study was limited in that the participants were a convenience sample recruited from one institution (see Vitta et al., 2022), and due to the small N-size, it lacked statistical power for meaningful tests of significance for score differences. Also, it is possible that initial exposure to target items may have primed meaning recall for later exposures. It would therefore be useful to conduct a similar study using a Latin Square or a counterbalanced design with a larger sample. Although more evidence is required before generalizations can be made, for this particular group of learners and these specific target words, the use of lenient marking on the full-sentence translation task of Test 3 was most similar to the criterion interview.

Acknowledgments

We would like to thank Mike McGuire and Dax Thomas for their helpful comments and suggestions on this paper. This study was partially funded by grant number JP 22K00793 from the Japan Society for the Promotion of Science.

References

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (Eds.). (2000). *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.). BMJ Books.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords.* Multilingual Matters. https://doi.org/10.21832/9781847690319

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Palgrave Macmillan. https://doi.org/10.1057/9780230293977

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

Stoeckel, T., Ishii, T., Kim, Y. A., Hung, T. H., Ho, N. T. P., & McLean, S. (2023). A comparison of contextualized and non-contextualized meaning-recall vocabulary test formats. *Research Methods in Applied Linguistics, 2*(3), 100075. https://doi.org/10.1016/j.rmal.2023.100075

Stoeckel, T., McLean, S., Raine, P., Ha, H. T., Ho., N. T. P., & Kim, Y. A. (2023). A contextualized meaning-recall vocabulary testing platform. *JALT Journal, 45*(2), 211–234. https://doi.org/10.37546/JALTJJ45.2-2

Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect-size driven sample size planning, randomization, and multi-site use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition, 44*(5), 1424–1448. https://doi.org/10.1017/S0272263121000541