THE JOURNAL OF TEACHING ENGLISH FOR SPECIFIC AND ACADEMIC PURPOSES
Vol. 5, Nº 4, 2017, pp. 719−726

UDC: 811.111’367.3:347.78.034 DOI: 10.22190/JTESAP1704719S

COMPARING MASTERY SENTENCE TEST SCORES WITH L2 TO L1 TRANSLATION TEST SCORES

Raymond Stubbe[1], Kousuke Nakashima[2]

[1]Nagasaki University, Japan
[2]National Institute of Technology, Ariake College, Japan

Abstract. *Mastery sentences, in which students compose a sentence demonstrating their understanding of a given English word, are recognized as an effective means of promoting vocabulary learning (Masson, 2012). As explained in Gallacher (2015, p.76) a "successful mastery sentence thus becomes one in which the target word, if removed, could only be replaced by a direct synonym." Early in the spring 2017 semester, students were advised that their vocabulary midterm test would be a Mastery Sentence test of 10 items. An explanation of Mastery Sentences was provided, as were successful and poor examples. High-beginner first year students, enrolled in a mandatory English class at a university in southern Japan (n = 209), took a Mastery Sentence midterm test of 10 items selected from their assigned vocabulary word list of 40 words. This test was given at the beginning of one class in June 2017. Towards the end of that same 90 minute class, students took an English to Japanese translation test of those same 10 individual items. Unfortunately, 81 students had perfect scores on the translation test, leading to a ceiling effect. These 81 were deleted from the data pool, leaving 128. Overall, mastery sentence test scores were higher than translation test scores. Results found that for 19% of the possible pairings, the mastery sentences did not match the translation; neither both were correct or both were incorrect. Also for half of the tested items more than 21% showed the same mismatch. It was concluded that mastery sentences did not consistently reflect actual word meaning knowledge.*

Key words: *mastery sentences, L2 to L1 translation test, vocabulary*

1. INTRODUCTION

Having students compose sentences (or longer pieces) to develop or demonstrate their vocabulary knowledge has been well researched in the past few decades (Folse, 2006; Hulstijn and Laufer, 2001; Kim, 2008; Laufer, 2003; Laufer & Hulstijn, 1998; Laufer & Paribakht, 1998; McNeil, 1996; Pichette, de Serres, & Lafontaine, 2011; Wesche and Paribakht, 1996; Webb, 2005, and others). Many of these studies support the use of sentence creation for vocabulary development, based on the notion of *depth of word processing* as originally proposed in Craik and Lockhart (1972). Laufer (2003), testing 10 low frequency words, reported that compared to a "reading group, the sentence writing group had significantly higher scores on the intermediate and the delayed (vocabulary) tests" (Laufer, 2003, p. 577). Hulstijn and Laufer (2001) similarly found that a composition task, incorporating all 10 target items, led to better passive recall scores on the tested items than either of the two reading tasks used in that study. Likewise, Pichette, de Serres, &

Submitted September 10th, 2017, accepted for publication October 30th, 2017

Lafontaine (2011) reported that their sentence writing group had superior recall of the tested words than their reading group on their immediate recall test. "However, delayed recall scores suggest that this superiority disappears over time" (Pichette, de Serres, & Lafontaine, 2011, p. 66). In Laufer (2003) learners wrote an original sentence for each target item. In Hulstijn and Laufer (2001) students wrote compositions incorporating the 10 targeted items. In and Pichette, de Serres, and Lafontaine (2011) students wrote three sentences for each item. Kim (2008) found that groups writing a composition versus writing original sentences led to "equal levels of performance for both groups in *initial* vocabulary learning and in longer-term *retention* of the target lexical items" (Kim, 2008, p. 310). Finally, Webb (2005) found that "when the allotted time on tasks depends on the amount of time needed for completion, with the (sentence) writing task requiring more time, the writing task was more effective" than the reading task (Webb, 2005, p. 33; 'sentence' added by the authors for clarification).

Not all vocabulary researchers agree that sentence writing is better than other vocabulary development activities. Folse (2006) found that that students retained words practiced under the three fill-in-the-blank exercises condition were acquired much better than those practiced under either of the other two exercise conditions examined in that study: single fill-in-the-blank exercises, and sentence writing exercises. "The findings suggest that the important feature of a given L2 vocabulary exercise is not depth of word processing but rather the number of word retrievals required" (Folse, 2006, p. 273).

In an earlier study, McNeill (1996) had Hong Kong English teachers complete a sentence production task, along with tests of word meaning for 30 lexical items. Results caused McNeill (1996, p. 39) to comment:

> Interestingly, the Hong Kong teachers' scores on a sentence production task were slightly higher than on tests of word meaning. This phenomenon suggests that the ability to produce convincing sentences in L2 may not be a reliable indicator of learners' understanding of the language produced.

McNeill is not alone in this finding. Both Paul Meara and John Read, discussing the *Vocabulary Knowledge Scale* (Paribakht & Wesche, 1993), a vocabulary test with a sentence writing component, warn against "accepting a student-composed sentence as evidence of word knowledge, despite the fact that it may be grammatically correct or semantically appropriate" (Read, 2000, p. 138). Additionally, according to Meara (1996, p.6):

> it is perfectly possible for learners to write sentences their [sic, should be 'that'] correctly illustrate the use of a particular word, even when they do not know the word's meaning. All they have to do is reproduce the context in which they first met the word, or reproduce a fixed expression which contains it.

This possibility of learners writing sentences that c*orrectly illustrate the use of a particular word, even when they do not know the word's meaning* is the focus of the present study.

2. Aim

The aim of study is to compare mastery sentence test scores with L2 to L1 translation test scores, to determine how well mastery sentences reflect actual word meaning knowledge. Mastery sentences, according to Masson (2012) are:

> elaborate sentences that indicate multi-level word understanding. Such sentences require the knowledge of a word be demonstrated beyond its spelling, and the meaning be explained within the sentence itself. A successful ‘mastery sentence’ must satisfy requirements for both usage and intended meaning in such a way that the target word can only be replaced by a direct synonym.

3. Methodology

At the beginning of the term students were advised that their midterm vocabulary test would be a mastery sentence test of 10 items, of the total 40 to be covered over the next 8 weeks. The concept of mastery sentences was explained, plus good and poor examples were given. Over the following 8 weeks, the vocabulary list of 40 words was presented (10 words bi-weekly) and students had to enter at least half in their vocabulary notebook (at least five entries bi-weekly). Each entry included the word; an L1 translation; and, their mastery sentence. Students were strongly encouraged to enter all words they did not know the meaning of.

A mastery sentence midterm was created that listed 10 selected words, randomly ordered, followed by blank lines where test-takers wrote their mastery sentences, as below:

confirm ____

The instructions for this test read “*For each word write a sentence that demonstrates the meaning of that word.”* Tests were given at the beginning of class, during regular class time, in June 2017. In all, five classes participated in this study. Participants were all high-beginner first-year students attending mandatory English classes at one Japanese university (n = 209).

An English to Japanese (L2 to L1) translation test was also created as a criterion measure to determine how well mastery sentences reflected student actual knowledge of the tested items. This decontextualized, single word translation test contained the same 10 English words, randomly ordered, and was given towards the end of the same class as the mastery sentence test. The L2 to L1 translation test format was selected as the criterion measure because: a) translation ability is a strong indicator of which words students can actually understand while reading (Waring & Takaki, 2003); b) meaning recall, which this style of translation test measures, "essentially corresponds to the lexical requirements of reading and listening (the word form is encountered and the meaning must be recalled)" (Pellicer-Sánchez & Schmitt, 2012, p. 494); and c) “asking participants to provide mother-tongue equivalents of the target language words was the most univocal way of verifying recognition” (Eyckmans, 2004, p. 77). Additionally, other leading vocabulary researchers agree that meaning recall L2 to L1 translation tests are an acceptable method of assessing the most important component of vocabulary knowledge, the form-meaning connection (Nation, 2001; Schmitt, 2010; Laufer & Goldstein, 2004; Nation & Webb, 2011). The popularity in Japan of the L2 to L1 translation test format as a criterion measure for other vocabulary tests can be seen in the following studies: Stubbe (2014); and Stubbe and Yokomitsu (2012); McLean, Stewart & Kramer, 2016; Stewart, McLean and Kramer (2016); Stoeckel & Stewart, 2016.

The mastery sentence test was marked by one of the authors, while the translation test was marked by the other author. Twenty of the translation test were randomly selected, copied and given to a different English teacher of Japanese descent for grading. Co-rater agreement on the 20 test forms was 96%. The same number of Mastery Sentence tests forms were similarly given to a native English speaker for marking. Co-rater agreement here was lower at 91%, likely because grading sentences require greater degrees of judgement.

4. Results

Table 1 presents the summary statistics for both the mastery sentence (MS) test and the translation (Tr) test. The mean score for the 209 participants on the MS test was higher than for the Tr test (9.1 vs. 8.8 out of ten; 1 mark for each item). The standard deviation (SD) for each test was quite low at 1.5 and 1.3, respectively. The range of scores was also quite similar. The difference between the two test means was statistically significant (t = 2.94, df = 208, p < .004), and the Cohen's d was .28 (Cohen, 1988 suggests .2 is a small effect size and .5 is a medium). Thus, although the difference between the means was statistically significant, it was still quite small. The correlation (Pearson product–moment) between the two test was not very strong (.48).

Table 1 Mastery sentence and translation test scores for 209 participants on 10 items

Statistic	MS test	Tr test
Mean	9.1	8.8
SD	1.5	1.3
Low	4	3
High	10	10
Correlation	.48	1

Note: Correlation on Tr test scores.

As informative as summary statistics are, they shed little light on how well each individual mastery sentence reflects actual student knowledge of that tested item. To accomplish this aim, a different analysis was undertaken. Each student's sentence score for each word in the MS test was compared with his or her answer to that same word on the Tr test. Total number of incidents and percentages for these matched forms are presented in Table 2. This direct comparison of MS and Tr test results resulted in a total of 2,090 comparisons (209 participants × 10 items). A correct sentence on the MS test matching with a correct answer on the Tr test had 1749 occurrences (83.7%, labeled *Correct – both tests*). An incorrect sentence on the MS test matched with a wrong answer on the Tr test occurred in only 89 instances (4.3%, *Incorrect – both tests*). Adding these categories together accounted for 1838 (87.9%) of total incidents (*Total Agreement).* In accordance with Meara (1996) and Read (2000), correct sentences were followed by incorrect translations on the Tr test had 156 instances (7.5%; labeled M.S. – O; Tr – X). The opposite (M.S. – X; Tr – O) occurred 96 times (4.3%).

Table 2 Mastery sentence & translation tests individual responses compared for 209 participants on 10 items.

Possible outcomes	# of Incidents	% of Incidents
Both tests – O	1749	83.7%
Both tests – X	89	4.3%
Total Agreement	1838	87.9%
MS test – O; Tr test – X	156	7.5%
MS test – X; Tr test – O	96	4.3%
Total Disagreement	243	11.6%

Note: O denotes correct, X denotes incorrect

A closer examination of the Tr test results revealed that 81 of the 209 participants had perfect scores (10 out of 10). This created a ceiling effect on this criterion measure, which could seriously skew further analyses because it would not allow the possibility of those students getting an MS test item correct while getting that same item incorrect on the Tr test. Accordingly, these 81students' MS and Tr results were deleted from the data pool, leaving 128 students. Table 3 presents summary statistics for both the mastery sentence (MS) test and the translation (Tr) test for these 128 participants. The difference in mean score for these 128 participants was even greater: 8.7 the MS test versus 8.1 on the Tr test. The standard deviation (SD) for each test was quite low at 1.5 and 1.3, respectively. The range of scores was also quite similar. The difference between the two test means was even more statistically significant ($t = 4.69$, $df = 127$, $p < .0001$) and the Cohen's *d* was larger at .42 (almost medium size). The correlation (Pearson product–moment) between the two test was even weaker (.4).

Table 3. Mastery sentence and translation test scores for 128 participants

Statistic	MS test	Tr test
Mean	8.7	8.1
SD	1.6	1.2
Low	4	3
High	10	9
Correlation	.40	1

Note: Correlation on Tr test scores.

As above, a second analysis comparing the two test forms was undertaken for these 128 participants. Total number of incidents and percentages for these matched forms are presented in Table 4. This direct comparison of MS and Tr test results resulted in a total of 1,280 comparisons (128 × 10 items). A correct sentence on the MS test matching with a correct answer on the Tr test had 962 occurrences (75.2%, down from the 83.7% of Table 2; same labels). An incorrect sentence on the MS test matched with a wrong answer on the Tr test occurred the same number of times (89), but the percentage has increased to from 4.3 to 7.0%. Adding these categories together accounted for 1051 (82.1%) of total incidents, down from 87.9%. Correct sentences followed by incorrect translations on the Tr test had 156 instances (12.2% up from 7.5%). The opposite (M.S. – X; Tr – O) remained at 96 occurrences (5.7%, up from 4.3). *Total Disagreement* went from 243 to 229, but increased in percentage from 11.6% to 17.9%.

Table 4 Mastery sentence & translation tests individual responses compared for 128 participants

Possible outcomes	# of Incidents	% of Incidents
Both tests – O	962	75.2%
Both tests – X	89	7.0%
Total Agreement	1051	82.1%
M.S. – O; Tr – X	156	12.2%
M.S. – X; Tr – O	73	5.7%
Total Disagreement	229	17.9%

Note: O denotes correct, X denotes incorrect

Table 4 (above) presents a comparison of both test forms for all 10 items. In a final *item* analysis, this comparative analysis was repeated for each individual tested item. As can be seen in Table 5, half of the 10 items had disagreement percentages between the two test forms (in bold) that were considerably higher than the other five (>21.0% versus <12.6%, respectively). The mean disagreement percentage for the first five words was 25.8% but only 10% for the final five words.

Table 5 Mastery sentence & translation test item responses compared for 128 participants.

Item	Both tests – O	Both tests – X	Total Agree	M.S. – O; Tr – X	M.S. – X; Tr – O	Total Disagree	Disagree %
Confirm	65	26	91	34	3	37	**28.9%**
Expression	82	9	91	26	11	37	**28.9%**
Envelope	73	19	92	26	10	36	**28.1%**
Key (adj.)	92	9	101	24	3	27	**21.1%**
Spotlight	100	0	100	15	13	28	**21.9%**
Package	110	2	112	9	7	16	12.5%
Carpenter	101	11	112	7	9	16	12.5%
Strategies	110	8	118	7	3	10	7.8%
Category	113	0	113	5	10	15	11.7%
Colleague	116	5	121	3	4	7	5.5%
Totals	962	73	156	89	1051	229	100%

Note: O denotes correct, X denotes incorrect

7. Conclusion

The aim of study was to compare mastery sentence test scores with L2 to L1 translation test scores, to determine how well mastery sentences reflect actual word meaning knowledge. Japanese first-year university students (n = 209) took a 10 item mastery sentence test followed by an L2 to L1 translation of the same items. Eighty-one test-takers (39%) scored ten out of ten on the translation test, creating a ceiling effect. These results were deleted from the data because they did not allow the possibility of a correct mastery sentence being followed by an incorrect translation, thereby skewing the comparison. Means for the remaining 128 test-takers were 8.7 and 8.1 for the mastery sentence and translation tests, respectively. Although the difference between the means was found to be statistically significant, the effect size was less than medium (.42). The correlation of .40

between the two tests was also not impressive, especially compared to the correlation of .67 reported in Pellicer-Sanchez and Schmitt (2012) between a yes-no test and a subsequent interview test, or the .72 reported in Stubbe (2014) between a yes-no test and an L2 to L1 translation test.

The analysis presented in Table 4 revealed that the mastery sentence test disagreed with the translation test in 19% of incidents. In over 12% of total instances a correct sentence was followed by an incorrect translation, meaning that the writer could produce an acceptable sentence without learning a proper meaning – as warned in Meara (1996) and Read (2000). In 7% of incidents the opposite occurred: students could not produce an acceptable sentence while producing a correct translation. The final item analysis (Table 5) revealed that for half of the tested words the disagreement rate between the two tests ranged from 21.1% to 29.8%. These figures, coupled with the low correlation (.4) between the two tests suggest that mastery sentences did not consistently reflect actual word meaning knowledge.

This study suffers from a number of weaknesses. As time did not allow for pre-testing student knowledge of the tested items, actual vocabulary acquisition cannot be measured. Additionally, the test items were generally too easy for this group of students with 38% of participants scoring perfectly on the criterion measure translation test, leading to a ceiling effect. Finally, the sample used in this study was a convenience sampling (intact university classes) and so the results cannot be generalized beyond these classrooms.

Future research should begin with a pre-test of a range of items to allow for measurement of actual knowledge gains. This pre-test should contain many more items than the planned mastery sentence and translation tests so the easy items can be eliminated to ensure the subsequent tests are not too easy thereby avoiding a ceiling effect.

REFERENCES

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

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour, 11,* 671–684.

Eyckmans, J. (2004). *Measuring receptive vocabulary size*. Utrecht, the Netherlands: LOT (Landelijke Onderzoekschool Taalwetenschap).

Folse, K. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273-293. doi: 10.2307/40264523

Gallacher, A. (2015). Mastery Sentences: A window into the interplay between word knowledge types. *Vocabulary Learning and Instruction 4* (1), 74-82.

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51,* 539–558.

Kim, Y. (2008). The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition. *Language Learning 58* (2), 285-325.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review/ LaRevue canadienne des langues vivantes, 59*(4), 567-587. doi: 10.3138/cmlr. 59.4.567.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength and computer adaptiveness. *Language Learning, 54*(3), 399–436. doi:10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22,* 1–26.

Laufer, B. & Paribakht, T.S., 1998. The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning, 48*(3), pp.365-391. doi: 10.1111/0023-8333.00046

McLean, S., Stewart, J. & Kramer, B. (2016). A Comparison of multiple-choice and yes/no test formats with a meaning-recall knowledge criterion. *Vocab@Tokyo Conference Handbook*, 127-128.

McNeil, A. (1996). Vocabulary knowledge profiles: evidence from Chinese-speaking ESL teachers. *Hong Kong Journal of Applied Linguistics 1* (1), 39–64.

Masson, M. (2012). Student feedback regarding the use of 'mastery sentences'. *Vocabulary Education and Research Bulletin 1* (1), 5-6.

Meara, P. (1996). The vocabulary knowledge frame work. Retrieved June 15, 2017 from: http://www.lognostics.co.uk/vlibrary/meara1996c.pdf

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge, UK: Cambridge University Press.

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Paribakht, T., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*, 9–29.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing, 29*(4), 489-509.

Pichette, F., de Serres, L., & Lafontaine, M. (2011). Sentence Reading and Writing for Second Language Vocabulary Acquisition. *Journal of Applied Linguistics, 32*(5), 1-18. doi: 10.1093/applin/amr037

Read, J. (2000). *Assessing Vocabulary.* Cambridge, UK: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* NY, NY: Palgrave Macmillan.

Stewart, J., McLean, S. & Kramer, B. (2016). Two empirical examinations of the effect of guessing on VST scores. *Vocab@Tokyo Conference Handbook*, 121-122.

Stoeckel, T. & Stewart, J. (2016). The "I don't know" option and L1 answer choices: A comparison of four variants of the Vocabulary Size Test. *Vocab@Tokyo Handbook*, 133-134.

Stubbe, R. (2014). Do Japanese students overestimate or underestimate their knowledge of English loanwords more than non-loanwords on yes-no vocabulary tests? *Vocabulary Learning and Instruction 3* (1), 29–43.

Stubbe, R. & Yokomitsu, H. (2012). English Loanwords in Japanese and the JACET8000. *Vocabulary Education & Research Bulletin, 1* (1), 10-11.

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.

Webb, S. (2005). Receptive and Productive Vocabulary Learning. *Studies in Second Language Acquisition, 27*(1), 33-52.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53*, 13–39.