SHIKEN

Volume 19 • Number 2 • November 2015

Contents

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 19 No. 2
November 2015

The Creation of a New Vocabulary Levels Test

Stuart McLean[1] and Brandon Kramer[2]
stuart93@me.com
1. Kansai University Graduate School
2. Momoyama Gakuin University

Abstract

This paper describes a new vocabulary levels test (NVLT) and the process by which it was written, piloted, and edited. The most commonly used Vocabulary Levels Test (VLT) (Nation, 1983, 1990; Schmitt, Schmitt, & Clapham, 2001), is limited by a few important factors: a) it does not contain a section which tests the first 1,000-word frequency level; b) the VLT was created from dated frequency lists which are not as representative as newer and larger corpora; and c) the VLT item format is problematic in that it does not support item independence (Culligan, 2015; Kamimoto, 2014) and requires time for some students to understand the directions. To address these issues, the NVLT was created, which can be used by teachers and researchers alike for both pedagogical and research-related purposes.

Keywords: vocabulary, assessment, levels, vocabulary levels test, vocabulary size

The purpose of this article is to provide a clear description of a new vocabulary levels test (NVLT) to assist teachers and researchers in its use. The NVLT was created as a parallel written receptive form of the Listening Vocabulary Levels Test (LVLT) (McLean, Kramer, & Beglar, 2015) and its creation therefore followed similar guidelines (see www.lvlt.info).

Vocabulary tests are often conceptualized as measuring either receptive or productive vocabulary knowledge, estimating either the total number of vocabulary items known (size tests) or mastery of vocabulary at certain frequencies of occurrence within a given corpus (levels tests). This paper introduces a new vocabulary levels test (NVLT), a receptive test of the most frequent 5,000 word families in Nation's (2012) British National Corpus / Corpus of Contemporary American English (BNC/COCA) word list. As the purposes and score interpretations of size and levels tests are often muddled within published research, the differences between the two types will be explained before describing the creation and intended interpretation of NVLT scores.

Measuring vocabulary size and interpreting vocabulary size test scores

Vocabulary size tests are intended to estimate the total number of words a learner knows. This estimate can be useful when comparing groups of learners, measuring long-term vocabulary growth, or providing "one kind of goal for learners of English as a second or foreign language" (Nation, 2013, p. 522). The Vocabulary Size Test (VST) (Nation & Beglar, 2007), for example, is a measure of written receptive word knowledge based on word family frequency estimates derived from the spoken subsection of the BNC (Nation, 2006). Each item on the VST presents the target word first in isolation followed by a non-defining context sentence, with four answer-choices presented in either English or in the learners' L1. Results of the VST among samples with a wide range in ability have shown that the test is able to reliably distinguish between learners of different vocabulary proficiency, either using the monolingual version (Beglar, 2010) or the various bilingual variants (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011).

Despite the VST's utility in separating students as a general measure of written receptive vocabulary knowledge *breadth*, inferences based on these results should be made with caution. For example, one of the stated interpretations of the VST is as an approximate estimate of known vocabulary. As the test samples 10 words each from the most frequent 1,000-word frequency bands (up to the 14th or 20th band depending on the version), "a test taker's score needs to be multiplied by 100 to get their total vocabulary size" (Nation, 2013, p. 525). A score of 30 out of 140, for example, would produce a size estimate of

3,000 known word families. While this score interpretation seems straightforward, it carries with it two assumptions which must be addressed: a) the target words on the VST are representative of the frequency bands which they were sampled from, so that each target word can be considered to represent 100 others, and b) correctly answering an item implies the written receptive knowledge of that target word. The first assumption, that the target words on the VST are representative of the frequency bands which they were sampled from, can be sufficiently assumed because the words were randomly sampled according to Nation and Beglar (2007). The second assumption, however, is a bit more problematic as the item format utilizes a 4-choice multiple-choice format, implying a 25% chance that the item would be correctly answered even if the examinee has absolutely no knowledge of the target word. While Nation (2012) recommends that all participants complete the entire 14,000-word version of the VST, McLean, Kramer, and Stewart (2015) showed that most correct answers for low proficiency students at the lowest frequency bands could be attributed to chance rather than lexical knowledge.

In order to increase the accuracy of the VST results, Beglar (2010), Elgort (2013), and McLean, Kramer, and Stewart (2015) recommend that students only take the test two levels above their ability. While this would reduce the previously mentioned score inflation due to mismatched items, the resultant score would not hold much pedagogical value. While some suggest that a VST score can be used to assign reading materials (Nation, 2013; Nguyen & Nation, 2011), this claim ignores the properties of the construct being measured (vocabulary *breadth*) as well as findings which argue that comprehension of reading materials require learners to know at least 95% of the words within the materials (e.g. Hsueh-chao & Nation, 2000; Laufer, 1989; van Zeeland & Schmitt, 2013). This is because while a vocabulary size score can give a rough estimate of the amount of words known, it does not imply knowledge of all vocabulary within that size estimate. For example, McLean, Hogg, and Kramer (2014) reported that the mean vocabulary size of Japanese university students (N = 3,427) was 3,396 word families (SD = 1,268) using the VST. These same learners, however, could not be said to have knowledge of the most frequent 3,396 word families, as all but the most able students had gaps in their knowledge of items from the first 1,000 words of English and all students failed to correctly answer multiple-choice items at the second and third 1,000-word bands.

Similar gaps have been found with the first and second 1,000-word frequency bands by Beglar (2010), Elgort (2013), Karami (2012), and Nguyen & Nation (2011). In order to measure knowledge of the most frequent vocabulary levels, a test made for that purpose is more appropriate.

Measuring knowledge of vocabulary levels and interpreting VLT scores

While the VST may be an appropriate instrument for separating students with a wide range of proficiencies, a more pedagogically useful measure of lexical knowledge is a test designed to measure the degree of mastery of the most frequent words of English. The most well-known of such tests, the Vocabulary Levels Test (VLT) (Nation, 1990; Schmitt, et al., 2001) was designed to provide richer information about learners' knowledge of the second, third, fifth, and tenth 1,000-word frequency bands, as well as Coxhead's (2000) Academic Word List (AWL). The primary purpose of a levels test such as this is to estimate learners' mastery of the most frequent vocabulary in the hope of assigning appropriate learning materials. For example, Nation (2013) states that meaning-focused reading input, which would include activities such as extensive reading and many kinds of task-based instruction, requires instructional materials to be written at a level with 95% known vocabulary. The test scores and their interpretations reflect this purpose, usually represented as a score out of 30 items for each level of the test, with mastery being a high proportion of correct answers at that level. Teachers can then use these results to help students focus on the most frequent unknown words until mastery is achieved.

Limitations of previous written vocabulary levels tests

While many have found the VLT (Nation, 1983, 1990; Schmitt, et al., 2001) useful in both pedagogy and research, Webb and Sasao (2013) identified a number of issues which this paper and the NVLT described within attempt to address.

Previous VLT content

The first limitation of the previous versions of the VLT is the absence of a section testing knowledge of the first 1,000-word frequency level, considered to be of the greatest value to learners because of the impact high frequency words have on comprehension. While the word families within the first 1,000-word frequency level account for 78% of the BNC corpus, the words from the next most frequent 1,000 word families account for only 8.1% (Nation, 2013).

Second, previous versions of the VLT sampled target words and distractors from outdated frequency lists. The first and second 1,000-word frequency levels used words from West's (1953) General Service List (GSL), and the 3,000, 5,000, and 10,000 word-frequency bands sampled words from lists constructed from Thorndike and Lorge (1944) and Kučera and Francis's (1967) frequency criteria. These lists represent the best effort to represent the language at the time they were made, but languages, and the vocabulary within them, are known to drift over time. In addition, advances in technology over the past few decades have allowed for improved corpus building, allowing researchers to collect much larger and more representative samples, which can be analyzed much more efficiently and accurately than the lists used to construct the VLT, allowing teachers to measure knowledge of vocabulary which would be considered much more appropriate for language learners today.

Previous VLT format

The previous VLT format (see Figure 1 for an example item cluster), which presents target items with distractors of the same vocabulary level, is problematic for several reasons: a) a lack of item independence, b) the relative inaccuracy of the format when compared with a standard four-choice item, c) student difficulty understanding the format, and d) difficulty adapting the tests to other testing mediums or base corpora.

1 business	
2 clock	_____ part of a house
3 horse	_____ animal with four legs
4 pencil	_____ something used for writing
5 shoe	
6 wall	

Figure 1. *Example of the VLT format.*

An assumption of test item analyses, whether within classical testing theory or item response theory (IRT), is that the items demonstrate what is called *item independence*. This means that the responses to different test items are not dependent on each other, meaning that they need to measure distinct aspects of knowledge. The VLT format (see Figure 1) displays six answer choices on the left, to be matched with the three target word definitions on the right. As students answer the three items, the number of available answer choices decreases, allowing them to answer more easily. Because of this, during their validation of VLT data, Beglar and Hunt (1999, p. 154) stated that "it has not been shown that the assumption of item independence holds true given this test format", a concern supported by Culligan (2015). Kamimoto (2014), looking into this issue specifically, concluded that the VLT format interacts with examinees'

knowledge of target items and causes local item dependence to various degrees and that this violation of item independence "comes from the test format" (p. 56).

Recently, Kremmel (2015) investigated the behavior of the different test formats in relation to qualitative interviews where the participants demonstrated knowledge of the target words. While both the item cluster VLT format and the standard multiple-choice format of the VST performed reasonably well, Kremmel found that the VLT format was slightly less representative of the participants' actual knowledge. This evidence suggests that the multiple-choice format more accurately measures vocabulary knowledge than the old levels test format, relative to the criterion of recall of meaning.

Previous use and piloting of the VLT format suggested that examinees may hesitate in answering VLT items and find its format problematic. The tests were piloted in a low English proficiency high school, and much time was necessary in order to carefully explain the testing procedure and allow the students to work through practice problems. In contrast, the standard multiple-choice format was immediately understood by the examinees, which facilitated a quicker administration of the test.

Finally, a standard multiple-choice format is also more easily adapted to online tests using widely available online testing software such as Survey Monkey <surveymonkey.com> or Moodle <moodle.org>, allowing teachers, researchers, and policy-makers to quickly administer and analyze tests or surveys with a large number of participants. A related limitation of this format it is that the distractors are not as easily edited as those within a standard multiple-choice item, as all distractors have to be considered in relation to the three target meanings. This would be particularly troublesome, for example, if a researcher tried to reorder the items to reflect a different wordlist which orders words differently, a problem further exacerbated if the lists utilize different word counting units.

The New Vocabulary Levels Test

In order to address the limitations stated above and provide an instrument with greater pedagogical utility, the authors created a new vocabulary levels test (NVLT). This NVLT is intended as a diagnostic and achievement instrument for pedagogical or research purposes, measuring knowledge of English lexis from the first five 1,000-word frequency levels of the BNC and the Academic Word List (AWL) (Coxhead, 2000). The test consists of five 24-item levels which together measure knowledge of the most frequent 5,000 word families, in addition to a thirty-item section which measures knowledge of the AWL. The entire 150-item test can be completed in 30 minutes; however, depending on the specific needs of researchers or teachers specific test sections can be administered in isolation.

NVLT format

The NVLT utilizes the multiple-choice format which provides multiple benefits: a) manipulation of distractor difficulty; b) efficient and reliable electronic marking; c) easily conducted item analyses; and d) item independence, a prerequisite for test analysis. Each item consists of four answer choices, from which examinees must select the word or phrase with the closest meaning to the target word. An example item is shown in Figure 2.

1. time: They have a lot of **time**.
a. money
b. food
c. hours
d. friends

Figure 2. An example item from the NVLT.

The piloted and revised test instructions (see Appendix A) are presently available in English and Japanese, with plans for additional languages in the future. When possible, to ensure that test instructions are understood, they should be given to examinees in their first language. To reduce the effects of guessing, the instructions state that if examinees have no knowledge of the correct answer, they should skip the question. However, if examinees feel that they may know the word, they should answer. The instructions also include two example questions to encourage understanding of the test format. Teachers and researchers should use the instructions they feel most appropriately meet their needs, while remembering that altering the instructions of the test may alter how items function.

The source of target vocabulary

The target words of the NVLT come from Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) word lists. The first and second 1,000-word family lists of the BNC/COCA were derived from a 10 million token corpus that consists of 6 million tokens from spoken British and American English. The corpus provides a list of high frequency words suitable for teaching and course design, and is a separate corpus than the one used to make the third to twenty-fifth 1,000-word family bands (Nation, 2012). The lists for the third 1,000-word family and above were created from BNC/COCA rankings once word families from the first 2,000 words of the BNC/COCA were removed. The BNC/COCA word lists include both British and North American varieties of English and are partly based on a spoken corpus, providing a strong basis for a monolingual vocabulary test (Nation, 2012). As Webb and Sasao (2013) stated, "the new BNC/COCA lists should be representative of current English and provide a far better indication of the vocabulary being used by native speakers today than the lists used for the creation of the earlier versions of the VLT" (p. 267).

The NVLT utilizes the word family unit because a) it was the unit utilized during the creation of the twenty-five 1,000-word BNC/COCA frequency lists (available with the Range software program, Heatley & Nation, 2015), b) even low proficiency learners have some control of word-building devices and they can perceive both a formal and semantic relationship between regularly affixed members of a word family (Nation & Beglar, 2007), c) it is consistent with the parallel LVLT and previous levels tests allowing for better comparison, and d) there is evidence that the word family is a psychologically real unit (Bertram, Baayen, & Schreuder, 2000; Bertram, Laine, & Virkkala, 2000; Nagy, Anderson, Schommer, Scott, & Stallman, 1989).

If given in its entirety the NVLT can measure knowledge of the first five 1,000-word frequency levels of the BNC/COCA and the AWL, which provides adequate coverage for numerous reading genres. As Webb and Sasao (2013) stated, "mastery of the 5,000 word level may be challenging for all but advanced learners, so assessing knowledge at the five most frequent levels may represent the greatest range in vocabulary learning for the majority of L2 learners" (p. 266).

Test creation

The items making up the first five 1,000-word frequency levels of the NVLT were created through a process of retrofit and redesign of previous Vocabulary Size Test (VST) items (Nation & Beglar, 2007). The previous validation of the use of the VST items with Japanese university students in an EFL context (Beglar, 2010) suggested their appropriateness to the NVLT which was piloted with a similar group. Item specifications (see Appendix B) were reverse engineered from previous test descriptions (e.g. Nation & Beglar, 2007) and specification-driven test assembly was implemented in line with Fulcher and Davidson (2007) when retrofitting items from three monolingual VST versions. Two VST versions were downloaded from <victoria.ac.nz/lals/about/staff/paul-nation> while the third version was obtained through personal correspondence with I.S.P. Nation. Items were re-assigned to their appropriate

BNC/COCA levels. For example, *period* and *basis* were relocated from the first 1,000-word level to the second 1,000-word level and items such as *nil*, present in the second 1,000-word frequency level of the VST, are not present in the NVLT as they do not occur in the first five 1,000-word levels of the BNC/COCA lists.

To ensure that the test is not conflating the construct of L2 contextual inferencing with vocabulary knowledge, the context sentences for each item were piloted using pseudowords in place of the target words. If the participants were then able to identify the correct answer without seeing the target word, the context sentence was edited as necessary.

The NVLT includes the AWL for three reasons: a) the importance of accessing AWL vocabulary knowledge because of the prominence of academic English programs; b) 10% coverage of tokens in academic texts is provided by the AWL (Coxhead, 2000); and c) previous tests measuring knowledge of the AWL have relied on the problematic VLT format.

AWL items were also created using the item specifications listed in Appendix B. The AWL is divided into nine 60-word and one 30-word sublists according to word frequency (Coxhead, 2000). Three target words were chosen from each of the first nine sublists and two from the tenth using a random number generator, and the final item was chosen at random from the entire AWL to ensure an even distribution of items. The final target word within each test item was the headword of the AWL word family (as listed in Coxhead, 2000). After each target word was chosen, distractor choices were randomly selected from the same sublist as the target word until the desired part of speech was obtained. If a suitable distractor could not be found in the same sublist, the process was repeated one sublist lower (i.e., the next higher frequency sublist).

Piloting was conducted to ensure that all distractors were plausible options. Then a generic sentence providing context without assisting the selection of the correct answer was written for each selected target item. Concordancer output from <www.lextutor.ca/conc/eng/> using the BNC/COCA corpus was consulted for authentic examples when the target word had numerous uses or meanings. When a sentence did not fit all of the distractors, the non-conforming distractor was replaced with randomly chosen words until all were found to fit the necessary criteria. Finally, each example sentence was checked to ensure that words from the first 1,000-word frequency level were used; however, a very limited number of words from the second 1,000 words of English were included, which were not found to be a problem in pilot testing. Repeated piloting of a small number of items continued until all significant problems were resolved.

Interpretability

Test interpretability is the degree to which qualitative meaning can be assigned to the quantitative measures produced by a test instrument (Medical Outcomes Trust Scientific Advisory Committee, 1995), and it is important for test creators to explicitly state how the test scores can be interpreted. The NVLT is intended as a test that measures an examinee's knowledge of the written form-meaning link of decontextualized vocabulary frequency bands. As a result, NVLT test scores should not be used to make statements about an examinee's productive vocabulary knowledge (see Laufer & Nation, 1999) or receptive aural vocabulary knowledge (see McLean, Kramer, & Beglar, 2015). It is recommended that the NVLT be utilized as a diagnostic, formative, or summative instrument, and that researchers and teachers use the 1,000-word frequency bands of the test that are appropriate for their needs. It is not recommended that the number of items per 1,000-word frequency level be reduced without careful IRT analysis.

While further research and testing is needed to empirically show the NVLT's utility in a variety of contexts, we can hypothesize potential uses for teachers and researchers. One example of an appropriate use of the

NVLT would be to assess learners' readiness for a particular course of study or the appropriateness of materials for learners. Instructors could first estimate the written vocabulary load of instructional materials or a single text. Given that research has shown that 98% coverage is ideal for easily comprehending written material (Hsueh-chao & Nation, 2000), the NVLT can be used to estimate learners' knowledge of lexis at particular word-frequency levels to determine whether they have the necessary lexical knowledge to comprehend course materials. For instance, learners who correctly answer at least 47-48 of the 48 items from the 1,000 and 2,000 word-frequency levels and half of the items from the 3,000 word-frequency levels on the NVLT would be deemed to have sufficient lexical knowledge to comprehend texts consisting of the most frequent 2,000 English word families. It should be remembered that this test is based on BNC/COCA word family lists. Thus, using the NVLT to assign level appropriate materials written based on different wordlists, and especially wordlists which use the lemma counting unit, is not recommended.

The NVLT could also be used to diagnose learners' vocabulary knowledge at the beginning of a course of study, estimate achievement throughout the course of study (i.e., formative assessment), and measure the knowledge gained upon completion of a course (i.e., summative achievement). For instance, if the goal of a beginner level course is to acquire knowledge of the 2,000 most frequent words of English, the threshold for mastering a single 1,000-word level should be at least 23 out of 24 correct items. Importantly, for higher frequency bands the necessity for a high mastery threshold is crucial, as any language user will commonly meet the highest frequency words when using the target language. This strict threshold is further supported by the mixed-methods validation of the aural version of this test (McLean, Kramer, & Beglar, 2015), which found that test-takers were more likely to correctly guess items that they did not know than to miss items that they knew. Similarly, mastery of the most frequent academic vocabulary should be defined as correctly answering 29 or more of the 30 AWL items.

Conclusion

The NVLT is a test that measures examinees' written receptive knowledge of the most frequent vocabulary frequency bands. The NVLT possesses four advantages over versions of the previous VLT: a) it measures vocabulary knowledge of each of the first five 1,000-word frequency bands; b) it measures vocabulary knowledge based on the more comprehensive and recent BNC/COCA; c) it utilizes a multiple-choice format facilitating item independence; and d) it has a parallel aural vocabulary levels test, the LVLT. It is recommended that the NVLT be used as a diagnostic, formative, or summative instrument, and that researchers and teachers utilize the 1,000-word frequency bands of the test that are appropriate for their needs. The test form is freely available and can be downloaded from <lvlt.info> or by contacting the authors.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101-118. doi: 10.1177/0265532209340194

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level vocabulary tests. *Language Testing*, *16*(2), 131-162. doi: 10.1177/026553229901600202

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*(3), 390-405. doi: 10.1006/jmla.1999.2681

Bertram, R., Laine, M., & Virkkala, M. M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, *41*(4), 287-296. doi: 10.1111/1467-9450.00201

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238. doi: 10.2307/3587951

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing, 32*(4), 503-520. doi: 10.1177/0265532215572268

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*(2), 253-272. doi: 10.1177/0265532212459028

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Heatley, A., & Nation, I. S. P. (2015). Range. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip

Hsueh-chao, M. H., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Kamimoto, T. (2014). Local item dependence on the Vocabulary Levels Test revisited. *Vocabulary Learning and Instruction, 3*(2), 56-68. doi: 10.7820/vli.v03.2.kamimoto

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal, 43*(1), 53-67. doi: 10.1177/0033688212439359

Kremmel, B. (2015). *The more, the merrier? Issues in measuring vocabulary size*. Paper presented at the LTRC 2015: The Language Testing Research Colloquium, Toronto.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press.

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*. Clevedon: Multilingual Matters.

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51. doi: 10.1191/026553299672614616

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the Vocabulary Size Test. *Vocabulary Learning and Instruction, 3*(2), 47-55. doi: 10.7820/vli.v03.2.mclean.et.al

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741-760. doi: 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*(1), 26-35. doi: 10.7820/vli.v04.1.mclean.et.al

Medical Outcomes Trust Scientific Advisory Committee. (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin, 3*(4), 1-4.

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 262-282. doi: 10.2307/747770

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1), 12-25. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1983-Testing-and-teaching.pdf

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, Mass.: Newbury House.

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

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved 17 September, 2012, from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13. Retrieved from http://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

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. doi: 10.1177/0033688210390264

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. doi: 10.1177/026553220101800103

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 Words*. New York: Teachers College Press, Columbia University.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457-479. doi: 10.1093/applin/ams074

Webb, S. A., & Sasao, Y. (2013). New directions In vocabulary testing. *RELC Journal, 44*(3), 263-277. doi: 10.1177/0033688213500582

West, M. P. (1953). *A general service list of English words*. London: Longman, Green & Co.

Appendix A

Translation of NVLT instructions

This is a vocabulary test.

Please select the option a, b, c or d which has the closest meaning to the word in **bold**.

Example question
see: They **saw** it.
a. cut
b. waited for
c. looked at The correct answer is **c**.
d. started

If you have no idea of the answer at all, please do not answer the question and move on to the next question.

However, if you think there is a chance that you may know the word, please try to answer.

Let's begin.

New Vocabulary Levels Test: 説明

これは単語力テストです。

太字になっている英語の意味に最も合う選択肢を a～d から選んでください。

問題例

see: They **saw** it.

a. cut

b. waited for

c. looked at 正解は **c** です。

d. started

答えが全く分からない場合は、空白にしておいてください。

しかし、わかる可能性があると思ったら、どうぞ挑戦してみてください。

では、始めましょう。

Appendix B

Specifications for New Items

Example Item:

school: This is a big **school**.
a. where money is kept
b. sea animal
c. place for learning
d. where people live

Overall

- The target word is presented in isolation and in bold within a context sentence
- The answer key should be randomly generated
- Avoid gender-biased language and have balanced gender representation

Target words

- Written in citation form
- From frequency list based on established corpus (BNC/COCA)
- Random sampling of words from each word-frequency level

Context sentence

- Context sentences in the first two 1,000-word levels should be written using vocabulary within the first 1,000-word level whenever possible
- Context sentences in the third 1,000-word level and above should be written using vocabulary within the first two 1,000-word levels whenever possible
- In cases where the part of speech is ambiguous, the most common form should be used based on frequency data
- The accompanying sentence should be as contextualized as possible without giving hints to the meaning of the target word

Distractors

- Core meanings of distractors should be of similar word frequency and difficulty level as the target word
- Distractors for items in the first two 1,000-word levels should be written using vocabulary within the first 1,000-word level whenever possible
- Distractors in the third 1,000-word level and above should be written using vocabulary within the first two 1,000-word levels whenever possible
- To as great a degree as possible, all distractors should be equally plausible in the context sentence

Minimal English Test: Item Analysis and Comparison with TOEIC Scores

Masaya Kanzaki
kanzaki-m@kanda.kuis.ac.jp
Kanda University of International Studies

Abstract

The Minimal English Test (MET) is a gap-filling dictation test developed by Maki, Wasada and Hashimoto (2003) with a view to evaluating the language proficiency of English learners quickly and easily. In this study, MET results were examined using item analysis to evaluate how well each item on the test functioned. Also, the results of the MET were compared with those of three different types of the Test of English for International Communication (TOEIC) in order to determine the degree to which the scores correlated. The participants in this study were 136 university students in Japan. They completed the MET and the TOEIC listening, reading and speaking tests. The MET results were analyzed for reliability and item statistics and the scores of the four tests were examined for correlations. The speaking test scores and MET scores correlated at .53, which is slightly higher than the correlation between the speaking test score and the scores of the listening and reading tests combined (r = .52).

Keywords: Minimal English Test, TOEIC, correlations, cloze testing, item analysis

Maki, Wasada and Hashimoto (2003) developed the Minimal English Test (MET) with the aim of creating a less expensive and more efficient alternative to commercially available English proficiency tests such as the Test of English for International Communication (TOEIC) and the Test of English as a Foreign Language (TOEFL). The MET was therefore designed to be administered quickly and easily; it takes a mere five minutes to complete and requires only two short passages with 72 blanks on a single A4-size sheet and an audio recording of the passages. The test-takers are required to listen to the audio and write down a word in each blank. The MET consists of passages with blanks and so it looks like a cloze test, first introduced by Taylor (1953) as a measure of the reading ability of native English speakers. Cloze tests, which require test-takers to fill in blanks with words to restore a text, have attracted a lot of attention from testing experts and teachers of English as a foreign language (EFL), and a number of research articles on these tests appeared in the literature from the 1960s to the 1980s (see Brown, 2013, for issues regarding cloze testing). There have been studies comparing cloze test scores with the results of the TOEFL (e.g. Darnell, 1968; Fotos, 1991; Irvine, Atai, & Oller, 1974) and other proficiency tests for EFL learners (e.g. Brown, 1988; Oller & Conrad, 1971; Stubbs & Tucker, 1974), many of which reported high correlation coefficients.

The MET, however, did not arise from this EFL tradition of cloze testing; it originated from a Japanese language test for non-native speakers of Japanese called the Simple Performance-Oriented Test (SPOT), developed by Kobayashi, Ford and Yamashita (1995). The SPOT consists of 60 unrelated sentences, each of which has one purposefully chosen *hiragana* character blanked out. Test-takers listen to an audio recording of the sentences and fill in the blanks, and completing the SPOT takes only a few minutes. Kobayashi et al. (1995) reported a correlation coefficient of .82 between the scores of the SPOT and Tsukuba University's placement test, a Japanese language proficiency test for students from overseas enrolled in the university, which consists of vocabulary, grammar, listening and reading sections and requires 150 minutes to complete. Goto, Maki and Kasai (2010, p. 95) called the MET "an English version of the SPOT" because it was modeled after the SPOT. The MET thus has three distinct features that are different from the majority of cloze tests. First, auditory cues are given to test-takers. Although a few cloze tests appearing in the literature had listening elements incorporated in them (e.g. Buck, 1988; Dickens & Williams, 1964; Henning, Gary, & Gary, 1983), providing auditory cues is not mainstream

practice for cloze testing. Second, the number of words between the two blanks in each line of the MET varies because blanked-out words are chosen according to word length (number of letters). In the creation of a typical cloze test, a fixed-rate deletion procedure (e.g., every twelfth word is blanked out) has been commonly used, although some studies have argued for rationally selecting words to delete instead of omitting them randomly at a fixed rate (e.g. Bachman, 1985; Brown, 1988). Third, the MET does not give test-takers much time during the test to stop and think about what words belong in the blanks; they have to proceed quickly from one blank to the next in order to keep up with the speed of the recording. Cloze tests, by contrast, usually allow ample time for test-takers to think about the content (e.g., 30 minutes to complete a 50-item cloze test based on a 400-word passage, as in Brown, 1988). In this respect, the MET is more of a word-recognition test than a cloze test.

Some correlation studies have been conducted to examine the relationships between the results of the MET and other English tests, such as the English test in the university entrance examination in Japan called the Center Test (with Goto, et al., 2010 reporting correlations ranging from $r = .60$ to $r = .72$) the TOEIC ($r = .74$, reported in Maki, Hasabe, & Umezawa, 2010), the STEP Eiken 2nd Grade ($r = .59$, reported in Maki & Hasabe, 2013), and the Vocabulary Levels Test ($r = .81$, reported in Kasai, Maki, & Niinuma, 2005). (For a list of papers on the MET, see Maki, 2015.) Kanzaki (2015a) compared the scores of the MET and the TOEIC listening, reading and speaking tests and obtained a correlation coefficient of .39 between the MET and the listening test, .51 between the MET and the reading test and .59 between the MET and the speaking test ($N = 90$). The present study is an expanded version of Kanzaki (2015a), with more participants and further analysis. In addition to comparing the results of the MET and three TOEIC tests for correlations, each item on the MET was analyzed using conventional item analysis in order to evaluate how well each item functioned.

Method

Data used in this study were collected over two years; first in July 2014, involving 90 participants, and second in July 2015, involving 46 participants. The MET and the listening, reading and speaking tests of the TOEIC were administered to the participants. Each item on the MET was analyzed for item statistics and the scores of the four tests were computed for correlations.

Participants

The study participants were 136 Japanese university students attending a private university specializing in foreign languages. They agreed to participate in the study in exchange for a cash reward of 1,000 yen, although they each had to pay the 3,500 yen to take the TOEIC listening and reading tests. The cost of the TOEIC speaking test was covered by a research grant. In 2014, 94 students signed up to take part in the study, but four of them were excluded because they scored 30 points or less on 72 questions of the MET; since they had left a lot of blanks unfilled, it was determined that they had not taken the test seriously. In 2015, 54 students signed up to take part, but four of them were excluded because they scored 30 points or less on the MET. Another four students, who had participated in the same study in the previous year, were excluded on the grounds that their MET scores might not be accurate since the same test was used in both years. The purposes of the study as well as the related procedures and requirements were explained to the participants before they signed a consent form.

Among the 136 participants, 10 were in their first academic year, 65 in their second, 29 in their third, and 32 in their fourth; 21 were male and 115 were female. In terms of fields of study, there were 76 international communication majors, 39 English language majors, 15 international business majors, two Chinese language majors, two Portuguese language majors, one Spanish language major and one Vietnamese language major. All the participants were native Japanese speakers except for two native

Korean speakers and one native Chinese speaker, who were fluent in Japanese. Three of the participants were enrolled in TOEIC-860 courses, eight in TOEIC-730 courses, 53 in TOEIC-650 courses and eight in TOEIC-600 courses (860, 730, 650 and 600 indicate the targeted TOEIC scores of these courses). The remaining 64 were not taking any TOEIC courses.

Materials

The MET and the TOEIC listening, reading and speaking tests were used in this study. The TOEIC listening and reading tests are always administered together and are therefore usually treated as two sections of one test. The TOEIC speaking test, on the other hand, can be taken independently when it is administered as the Institutional Program (IP), with which each institution sets the time, date and place of the exam. The three TOEIC tests used in the study were administered as IP tests.

Minimal English Test (MET).

The MET consists of two passages, one with 200 words and the other with 198 words. Both of them are taken from an English textbook for university students written by Kawana and Walker (2002). The audio recording that accompanies the textbook is also used for the MET. The two passages are spread out over 36 lines of between 6 and 17 words each, and the average number of words per line is 11. Each line has two blanks, and only words that have four letters or fewer have been blanked out, because such short words are considered to be the English equivalent of one *hiragana* character deleted in the SPOT, after which the MET was modeled. Because of this restriction, the deletion frequency of the MET is irregular; the number of words between two blanks ranges from 0 to 10 (4.24 on average), excluding the interval between the last blank of the first passage and the first blank of the second passage, which has 15 words. (For the actual test sheet, with item numbers and an answer key, see the Appendix.) Test-takers listen to the passages, recorded at a rate of about 125 words per minute, and fill in 72 blanks. There is a 10-second pause between the two passages (between lines 18 and 19). The test ends as soon as the audio recording finishes and no extra time is provided for going back to fill in any remaining blanks; therefore, test-takers have to write down words quickly and keep up with the speed of the recording. Because auditory cues are given, the exact word scoring procedure (only the intended word is accepted as the correct answer) is used in the marking of the test, and spelling mistakes are counted as wrong answers. However, the author of this paper made one exception for the misspelling of paid in line 9, #17, such as *payed*, *peid* and *paied*, on the grounds that those who misspelled the word in such ways were able to hear it correctly and knew that it was the past form of *pay*.

TOEIC Listening Test.

The TOEIC listening test consists of 100 multiple-choice questions, and raw scores of between 0 and 100 are converted to scaled scores of between 5 and 495. The test has four parts, the details of which are shown in Table 1.

Table 1
Four Parts of the TOEIC Listening Test

Part	Task	# of Qs
1	For each question with a photo, listen to four sentences and choose the one that best describes the image.	10
2	Listen to a question or statement followed by three responses and choose the most appropriate response.	30
3	Listen to a conversation and answer comprehension questions.	30
4	Listen to a short talk and answer comprehension questions.	30

TOEIC Reading Test.

The TOEIC reading test consists of 100 multiple-choice questions, and raw scores of between 0 and 100 are converted to scaled scores of between 5 and 495. The test has three parts, the details of which are shown in Table 2.

Table 2
Three Parts of the TOEIC Reading Test

Part	Task	# of Qs
5	Choose a word or phrase to fill in a blank in a sentence.	40
6	Choose words or phrases to fill in blanks in a passage.	12
7	Read a passage or a set of two passages and answer comprehension questions.	48

Note. The TOEIC reading test starts with Part 5 because it immediately follows the TOEIC listening test, which ends with Part 4, and the two tests are always taken as a set.

TOEIC Speaking Test.

The TOEIC speaking test is a computer-based examination requiring test-takers to sit in front of a computer while wearing a headset with a microphone. Instructions are provided on the computer screen and through the headset. Test-takers speak into the microphone and their speeches are recorded and sent to certified raters for evaluation. There are 11 questions in the test and scores are given in the range of 0 to 200. Table 3 shows the details of the test.

Table 3
Details of the TOEIC Speaking Test

Question #	Task
1–2	Read aloud the text that appears on the screen.
3	Describe the picture on the screen.
4–6	Answer three questions about a single topic as though you are participating in a telephone interview.
7–9	Read the information on the screen and answer three questions about it as though you are responding to a telephone inquiry.
10	Listen to a recorded message about a problem and propose a solution for it.
11	Express an opinion about a specific topic.

Procedure

Both of the data collection sessions, one in July 2014 and the other in July 2015, took place on campus over two days. The participants took the TOEIC listening and reading tests on the first day and the MET and TOEIC speaking test on the second day. The author of this paper marked the MET and the results were entered into a Microsoft Excel sheet and then used for item analysis. The results of the three TOEIC tests were provided by the Institute for International Business Communication, the administrator of the TOEIC in Japan. The scores of the four tests were compared for correlations.

Analysis

First, descriptive statistics of the four tests, such as means, standard deviations and minimum and maximum scores, were calculated. Second, the reliability index (Cronbach's alpha) and the standard error of measurement (SEM) were computed. Reliability indices and the SEM for the three TOEIC tests could not be calculated because the Educational Testing Service (ETS), the developer and administrator of the

TOEIC, discloses neither the item-by-item results nor raw scores. Third, each item on the MET was analyzed for item facility and item discrimination. Finally, the scores of the four tests were compared for correlations. Descriptive statistics and correlations were computed using IBM SPSS Statistics for Windows (2013) and the calculation of the reliability index and SEM as well as item analysis were carried out using Microsoft Excel (2013).

Results and Discussion

Descriptive Statistics

Table 4 shows the descriptive statistics for the scores of the MET and three TOEIC tests. The participants performed better on the TOEIC listening test than on the TOEIC reading tests, as the average listening test score was 102.39 points higher than the average reading test score. The average combined score of the listening and reading tests was 649.30 (ranging from 310 to 945) with a standard deviation of 120.01.

Table 4
Descriptive Statistics for the MET and Three TOEIC Tests (N = 136)

Test	Score Range	Mean	SD	Minimum	Maximum
MET	0–72	47.84	8.77	31	70
TL	5–495	375.85	56.67	170	495
TR	5–495	273.46	74.85	100	475
TLR	10–990	649.30	120.01	310	945
TS	0–200	118.31	21.35	60	180

Note. TL = TOEIC listening test, TR = TOEIC reading test, TLR = TOEIC listening and reading tests combined, TS = TOEIC speaking test.

Reliability and Standard Error of Measurement

The reliability index for the MET was .86 and the SEM based on Cronbach's alpha was 3.3, which means that if the same person were to take the MET repeatedly, his or her score would be within the range of plus or minus 3.3 of the current score 68% of the time.

Reliability estimates for the three TOEIC tests used in this study could not be calculated since the ETS does not make item-by-item results or raw scores available. However, the ETS (Educational Testing Service, 2013, p. 16) reported that the reliability index (KR-20) of the TOEIC listening and reading scores across all forms of their norming samples is "approximately .90". Also, the ETS (Educational Testing Service, 2010, p. 18) reported that the reliability of the TOEIC speaking test is .80 "based on the data from January 2008 to December 2009 administrations in the Public Testing Program". The reliability estimate of the same test, however, differs when it is taken by a different group of test-takers, and therefore the estimates for the three tests taken by the participants of this study may not be the same as the aforementioned figures reported by the ETS. They are probably lower than the ETS figures because the sample size of this study is much smaller.

Similarly, the SEM for the three TOEIC tests taken by the participants of this study cannot be calculated, but the ETS (Educational Testing Service, 2013, p. 16) reported that the SEM is "about 25 scaled score points" for each of the TOEIC listening and reading tests. The ETS (Educational Testing Service, 2010, p. 18) also reported that "based on the same datasets used for reliability estimates, the SEM is approximately 13 scale points" for the TOEIC speaking test.

Item Analysis

Table 5 shows the item facility and discrimination indices for the 72 items on the MET. Item facility, which is sometime called "item difficulty," indicates the percentage of participants who answered a particular item correctly. It can be obtained by dividing the number of the participants who answered a certain item correctly by the total number of participants (Brown, 2005). Item discrimination indicates how well a certain item discriminates those who performed well on the test as a whole from those who did not. In this study, a point-biserial correlation is used as an item discrimination index. This is a correlation between the results of an individual item and the total test scores and can be obtained by "comparing a dichotomous nominal scale (the correct or incorrect answer on each item usually coded as 1 or 0) with a continuous scale (total scores on the test)" (Brown, 2005, p. 162).

Table 5
Item Statistics for the MET (N = 136)

Item	*IF*	*ID*	Item	*IF*	*ID*	Item	*IF*	*ID*
1	.88	.13	25	.93	.19	49	.76	.36
2	.85	.21	26	.88	.10	50	.70	.48
3	.99	.06	27	.41	.48	51	.60	.39
4	.90	.18	28	.58	.36	52	.94	.14
5	.44	.25	29	.61	.44	53	.91	.24
6	.54	.30	30	.82	.34	54	.24	.50
7	.93	.00	31	.83	.40	55	.76	.31
8	.93	.19	32	.97	.20	56	.41	.49
9	.94	.28	33	.71	.25	57	.21	.38
10	.79	.36	34	.52	.42	58	.40	.49
11	.95	.18	35	.69	.38	59	.36	.35
12	.60	.28	36	.35	.42	60	.28	.25
13	.32	.42	37	.92	.22	61	.65	.34
14	.85	.29	38	.41	.39	62	.40	.36
15	.84	.41	39	.40	.26	63	.61	.12
16	.96	.17	40	.90	.10	64	.29	.33
17	.74	.20	41	.66	.29	65	.46	.26
18	.13	.37	42	.97	.22	66	.07	.25
19	.63	.44	43	.92	.24	67	.74	.22
20	.93	.23	44	.94	.06	68	.82	.40
21	.91	.32	45	.18	.34	69	.49	.39
22	.76	.47	46	.77	.25	70	.76	.28
23	.30	.39	47	.85	.17	71	.36	.45
24	.95	.17	48	.79	.29	72	.53	.25

Note. IF = item facility, *ID* = item discrimination (point biserial).

The item facility indices ranged from .07 to .99 with the mean of .66. The standard deviation was .25, which indicates the average distance from the mean. This shows that the item facility indices dispersed fairly widely, indicating that the difficulty levels of the items varied widely since item facility indices show how easy each item is (the higher the number, the easier).

The item discrimination indices (point biserial) ranged from .00 to .50 with the mean of .29. Ebel (1979, as cited in Brown, 2005) suggested that an item discrimination index of .40 or higher indicates that the item is "very good" in terms of separating the high and low achieving participants, between .30 and .39 is "reasonably good," between .20 and .29 is "marginal," and below .19 is "poor." According to this guideline, 15 items out of 72 were "very good," 19 were "reasonably good," 23 were "marginal," and 15 were "poor." Among the 15 items with an item discrimination index of .19 or lower, 11 had an item facility index of .90 or higher. It seems that those items were too easy to separate the high and low achieving participants. The item facility indices of the items with a discrimination index of .40 or higher ranged from .24 to .84, and 10 of them were between .30 and .70. Although the item with the highest discrimination index, #54, has an item facility index of .24, those items with the middle-range facility indices tend to have high discrimination indices.

Correlations

TOEIC Tests

Table 6 shows the correlations between the scores of the three TOEIC tests. Among the three combinations, the highest correlation was between the listening and reading test scores (r = .66) and the lowest was between the listening and speaking test scores (r = .46). Between these was the correlation between the reading and speaking test scores (r = .48). This order is unusual when compared to findings reported in other correlation studies involving the three TOEIC tests, as the correlation between the listening and speaking test scores is usually higher than the correlation between the reading and speaking test scores. For example, Liao, Qu and Morgan (2010) reported a correlation of .76 between the listening and reading test scores, .66 between the listening and speaking test scores, and .57 between the reading and speaking test scores. Liu and Constanzo (2013) reported .73, .63 and .54, and Kanzaki (Kanzaki, 2015b) reported .68, .50 and .48 in the same order. The lower correlations between the listening and speaking test scores and between the reading and speaking test scores could be due to the lower reliability of the speaking test. The reliability of the speaking test reported by the Educational Testing Service (2010, p. 18) is .80, whereas the reported reliabilities of the listening and reading tests (Educational Testing Service, 2013, p. 16) are "approximately .90". It has been pointed out in the literature that when reliability estimates are low, the correlations will likely be underestimated (e.g. Ayearst & Bagby, 2011; Spearman, 1904).

Table 6
Correlations between the Three TOEIC Tests (N = 136)

	TL	TR	TS
TL	1.00	.66*	.46*
TR		1.00	.48*
TS			1.00

Note. TL = TOEIC listening test, TR = TOEIC reading test, TS = TOEIC speaking test.
* = $p < .001$

The correlation between the speaking test score and the combined score of the listening and reading tests was .52 ($p < .001$).

MET versus TOEIC

Table 7 shows the simple and disattenuated correlations between the MET and the three TOEIC tests. The second row of the table shows the simple correlations. The MET score correlated with the speaking test score at .53 and the figure was higher than those between the MET and the listening test (r = .38) and the

MET and the reading test (r = .48). The MET score correlated with the combined score of the listening and reading tests at .48, which is lower than the correlation reported by Maki et al. (2010) (r = .74, N = 57).

Table 7
Simple and Disattenuated Correlations between the MET and TOEIC (N = 136)

	TL	TR	TLR	TS
Simple *r* with MET	.38*	.48*	.48*	.53*
Disattenuated *r* with MET	.44*	.54*	.54*	.64*

Note. TL = TOEIC listening test, TR = TOEIC reading test, TLR = TOEIC listening and reading tests combined, TS = TOEIC speaking test.
* = $p < .001$

The third row of Table 7 shows the disattenuated correlations, figures that have been corrected for attenuation. Spearman (1904) noted that raw correlations are lower than true correlations because of measurement errors and, therefore, in order to estimate the real correlations, the raw figures need to be corrected on the basis of reliability estimates. He suggested the following equation for correction for attenuation:

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

where

r_{xy} = correlation between x and y
r'_{xy} = correlation between x and y, corrected for attenuation
r_{xx} = reliability of x
r_{yy} = reliability of y

(Adapted from Murphy & Davidshofer, 2004, p. 137)

The formula requires reliability estimates. However, since the ETS does not disclose the reliabilities of the scores of the TOEIC tests taken in this study, they remain unknown. I used the figures reported in the ETS publications (.90 for the listening and reading tests and .80 for the speaking test) instead, assuming that the reliability estimates of this particular group were lower than those reported by the ETS due to a smaller sample size and, therefore, using the ETS figures would provide conservative estimates of true correlations. Along with these ETS figures, a reliability estimate of .86 for the MET was used in correction for attenuation. When disattenuated, the correlations with the listening, reading and combined listening and reading scores increased by .06 each, whereas the correlations with the speaking score increased by .11, making the MET's closer relationship with the speaking score more distinct.

It is surprising that the correlation between the MET and listening test was the lowest among the four combinations, considering that the MET contains listening elements. Moreover, ordinary cloze tests, which do not provide auditory cues, "have consistently correlated best with measures of listening comprehension" (Oller, 1973, p. 114). Unexpectedly, a test with listening elements correlated poorly with a listening test, while the tests that did not have listening elements correlated well with measures of listening comprehension. One possible explanation for this is since the participants, whose average TOEIC listening test score was 375.85 out of 495, easily understood the recorded text for the MET, the test did not function as a tool for measuring listening abilities.

Another surprise was that the correlation between the MET and speaking test scores was the highest among the four combinations. The MET does not test speaking skills directly; however, it seems that the

test tapped the speaking abilities of the participants. One characteristic of the MET that might relate to speaking abilities is the test's multitasked nature. Test-takers listen to the audio recording, read the text, write down words and, at the same time, anticipate what will come next. Similar multitasked abilities are needed for speaking. In addition, test-takers have to move quickly from one blank to the next in order to perform well on the MET. This quickness is also necessary for the TOEIC speaking test, for which test-takers have to complete tasks within a given timeframe and the time pressure is higher than in the TOEIC listening and reading tests.

Conclusion

The reliability index of the MET was .86, which indicates that the MET scores in this study were fairly reliable. The results of the item analysis on the MET revealed that 15 items out of 72 did not function well in separating the high and low achieving participants. The quality of the test might be improved by eliminating these poorly functioning items. It would be interesting to see how such a revision would affect the reliability of the test and correlations with the TOEIC.

The correlation between the TOEIC listening and speaking scores was .46, which was lower than the correlation between the reading and speaking test scores (r = .48). Logically, a speaking test, which deals with spoken English, should have a closer relationship with a listening test measuring the receptive skill of spoken English than a reading test measuring the receptive skill of written English, but the scores of the three TOEIC tests in this study did not reflect that logic.

Among the three TOEIC tests, the MET most strongly correlated with the speaking test (r = .53, and .64 after disattenuation) and most poorly with the listening test (r = .38, and .44 after disattenuation). It seems that the MET did not measure listening abilities even though the test-takers had listened to the audio recording during the test. The correlation between the speaking test score and the combined score for the listening and reading tests was .52, which suggests that the MET can be as good a predictor of speaking abilities as the TOEIC listening and reading tests, although a raw correlation of .53 and a disattenuated correlation of .64 between the MET and speaking test scores is not high enough to replace the speaking test with the MET for the purpose of measuring speaking abilities.

Acknowledgements

The author is grateful to Dr. Hideki Maki of Gifu University for providing the MET-related materials. Also, the author would like to thank Professor Norihito Kawana of Sapporo International University and Seibido Shuppan Cooperation for granting permission to reproduce the MET in this paper. This study was supported by JSPS KAKENHI Grant Number 25370727.

References

Ayearst, L. E., & Bagby, R. M. (2011). Evaluating the psychometric properties of psychological measures. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of Assessment and Treatment Planning for Psychological Disorders* (2nd ed., pp. 23-61). New York, NY: Guilford Press.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*(3), 535-556. doi: 10.2307/3586277

Brown, J. D. (1988). Tailored cloze: improved with classical item analysis techniques. *Language Testing, 5*(1), 19-31. doi: 10.1177/026553228800500102

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New ed.). New York: McGraw-Hill.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies, 7*(1), 1-32. Retrieved from http://www.researchgate.net/publication/283443076_IJLS_Journal_7%281%29_January_2013_Full_Text

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal, 15*(1), 15-42.

Darnell, D. K. (1968). The development of an English language proficiency test of foreign students, using a clozentropy procedure: Final report. Boulder, CO: University of Colorado, US DHEW Project No. 7-H-010, ERIC ED 024039. Retrieved from http://files.eric.ed.gov/fulltext/ED024039.pdf

Dickens, M., & Williams, F. (1964). An experimental application of "cloze" procedure and attitude measures to listening comprehension. *Speech Monographs, 31*(2), 103-108. doi: 10.1080/03637756409375397

Ebel, R. L. (1979). *Essentials of educational measurement*. Englewood Cliff, NJ: Prentice-Hall.

Educational Testing Service. (2010). TOEIC user guide: Speaking and writing. Retrieved from http://www.ets.org/s/toeic/pdf/toeic_sw_score_user_guide.pdf

Educational Testing Service. (2013). TOEIC user guide: Listening & reading. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning, 41*(3), 313-336. doi: 10.1111/j.1467-1770.1991.tb00609.x

Goto, K., Maki, H., & Kasai, C. (2010). The Minimal English Test: A new method to measure English as a Second Language proficiency. *Evaluation & Research in Education, 23*(2), 91-104. doi: 10.1080/09500791003734670

Henning, G., Gary, N., & Gary, J. O. (1983). Listening recall: A listening comprehension test for low proficiency learners. *System, 11*(3), 287-293. doi: 10.1016/0346-251X(83)90046-5

Irvine, P., Atai, P., & Oller, J. W. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24*(2), 245-252. doi: 10.1111/j.1467-1770.1974.tb00506.x

Kanzaki, M. (2015a). *Minimal English Test vs. three TOEIC tests*. Paper presented at the JALT PanSIG2015, Kobe, Japan.

Kanzaki, M. (2015b). TOEIC survey: Speaking vs. listening and reading. In P. Clements, A. Krause & H. Brown (Eds.), *JALT2014 Conference Proceedings* (pp. 639-649). Tokyo, Japan: JALT.

Kasai, C., Maki, H., & Niinuma, F. (2005). The Minimal English Test: A strong correlation with the Paul Nation Proficiency Test. *岐阜大学地域科学部研究報告 (Bulletin of the Faculty of Regional Studies, Gifu University), 17*, 45-52. Retrieved from https://repository.lib.gifu-u.ac.jp/bitstream/123456789/4588/1/KJ00004182420.pdf

Kawana, N., & Walker, S. (2002). *This is media.com*. Tokyo: Seibido.

Kobayashi, N., Ford, J., & Yamamoto, H. (1995). 日本語能力簡易試験(SPOT)の得点分布傾向：中上級向けテストと初級向けテスト (Distribution of Scores in the Simple Performance-Oriented

Test (SPOT): comparison of scores between easy and difficult versions of SPOT). *筑波大学留学生センター日本語教育論集 (Journal for Japanese Language Education, University of Tsukuba), 10*, 107-119.

Liao, C. W., Qu, Y., & Morgan, R. (2010). The relationships of test scores measured by the TOEIC Listening and Reading Test and TOEIC Speaking and Writing Tests. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/TC-10-13.pdf

Liu, J., & Costanzo, K. (2013). The relationship among TOEIC listening, reading, speaking, and writing skills. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/TC2-02.pdf

Maki, H. (2015). The Minimal English Test (MET) Retrieved 2 November, 2015, from http://www.geocities.jp/makibelfast/MET.html

Maki, H., & Hasabe, M. (2013). The Minimal English Test and the Test in Practical English Proficiency by the STEP. *岐阜大学地域科学部研究報告 (Bulletin of the Faculty of Regional Studies, Gifu University), 32*, 25-30. Retrieved from http://repository.lib.gifu-u.ac.jp/bitstream/123456789/45802/1/reg_030032003.pdf

Maki, H., Hasabe, M., & Umezawa, T. (2010). A study of correlation between the scores on the Minimal English Test (MET) and the scores on the Test of English for International Communication (TOEIC). *岐阜大学地域科学部研究報告 (Bulletin of the Faculty of Regional Studies, Gifu University), 27*, 53-63. Retrieved from http://repository.lib.gifu-u.ac.jp/bitstream/123456789/34847/1/reg_030027005.pdf

Maki, H., Waseda, H., & Hashimoto, E. (2003). Saishoo Eego Tesuto: Shoki kenkyuu (The Minimal English Test: A preliminary study). *Eego Kyooiku (The English Teachers' Magazine) 53*(10), 47-50.

Murphy, K. R., & Davidshofer, C. O. (2004). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, N.J.: Pearson Education.

Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*(1), 105-118. doi: 10.1111/j.1467-1770.1973.tb00100.x

Oller, J. W., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*(2), 183-194. doi: 10.1111/j.1467-1770.1971.tb00057.x

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101. doi: 10.2307/1412159

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *The Modern Language Journal, 58*(5-6), 239-241. doi: 10.1111/j.1540-4781.1974.tb05105.x

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Appendix

MET with item numbers and answer key

1. The majority of people have at least one pet at (1. some) time in their (2. life).
2. Sometimes the relationship between a pet (3. dog) or cat and its owner is (4. so) close
3. that (5. they) begin to resemble (6. each) other in their appearance and behavior.
4. On the other (7. hand), owners of unusual pets (8. such) as tigers or snakes
5. sometimes (9. have) to protect themselves (10. from) their own pets.
6. Thirty years (11. ago) the idea of an inanimate (12. pet) first arose.
7. This was the pet (13. rock), which became a craze (14. in) the United States and
8. spread (15. to) other countries as (16. well).
9. People (17. paid) large sums of money for ordinary rocks and assigned (18. them) names.
10. They tied a leash around the rock and pulled (19. it) down the street just (20. like) a dog.
11. The rock owners (21. even) talked (22. to) their pet rocks.
12. Now (23. that) we have entered the computer age, (24. we) have virtual pets.
13. The Japanese Tamagotchi—(25. the) imaginary chicken (26. egg)—
14. (27. was) the precursor of (28. many) virtual pets.
15. Now there (29. are) an ever-increasing number of such virtual (30. pets)
16. which mostly young people are adopting (31. as) their (32. own).
17. And (33. if) your virtual pet (34. dies),
18. you (35. can) reserve a permanent resting place (36. on) the Internet in a virtual pet cemetery.
19. Sports are big business. Whereas Babe Ruth, the (37. most) famous athlete of (38. his) day,
20. was well-known (39. for) earning as (40. much) as the President of the United States, the average
21. salary (41. of) today's professional baseball players is (42. ten) times that of the President.
22. (43. And) a handful of sports superstars earn 100 times (44. more) through their contracts
23. (45. with) manufacturers of clothing, (46. food), and sports equipment.
24. But every generation produces (47. one) or two legendary athletes (48. who) rewrite
25. the record books, and whose ability and achievements (49. are) remembered (50. for) generations.
26. (51. In) the current generation Tiger Woods and Michael Jordan are two (52. such) legendary
27. figures, (53. both) of whom (54. have) achieved almost mythical status.
28. The (55. fact) that a large number of professional athletes (56. earn) huge incomes
29. has (57. led) to increased competition throughout (58. the) sports world.
30. Parents (59. send) their children to sports training camps (60. at) an early age.
31. Such (61. kids) typically practice three to (62. four) hours a day,
32. (63. all) weekend (64. and) during their school vacations
33. in order (65. to) better their chances of eventually obtaining (66. a) well-paid position
34. on a professional (67. team) when they grow (68. up).
35. As for the (69. many) young aspirants who do (70. not) succeed,
36. one wonders if they (71. will) regret having (72. lost) their childhood.

Questions and answers about language testing statistics:

Characteristics of sound quantitative research

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

QUESTION:

In Brown, 2005, you explained the characteristics of well-done qualitative research by explaining the importance of dependability, credibility, confirmability, and transferability. You mentioned in passing that the parallel characteristics for quantitative research were reliability, validity, replicability, and generalizability. But you never really explained those quantitative research characteristics. I think it would be useful to know more about those characteristics of sound quantitative research and maybe even something about the characteristics of good quality mixed-methods research. Could you talk about these other research paradigms?

ANSWER:

Certainly, let me begin by reviewing my definition of what I think research is. Then I will turn to the issues that quantitative researchers need to address in order to produce sound quantitative research by explaining four concepts: reliability, validity, replicability, and generalizability. As I proceed through these explanations, you will see how similar and yet different the qualitative and quantitative sets of characteristics are. I will focus on the characteristics of quantitative research here and save the characteristics of mixed-methods research for a subsequent column (Brown, forthcoming in 2016).

What is research?

In the column you refer to (Brown, 2005), I defined research very broadly as: "any systematic and principled inquiry" (based on Brown, 1992, 2004). Research can be systematic and principled in many different ways. As I discussed in Brown (2005), sound qualitative research (at one end of the continuum) can be systematic in terms of its dependability, credibility, confirmability, and transferability, while sound quantitative research can be systematic in terms of its reliability, validity, replicability, and generalizability—four characteristics that will serve as the focus of the rest of this column.

Reliability

In quantitative research, at a micro level, reliability can be defined something like the degree to which the results of research measurements and observations are consistent. The reliability of a study's measurements and observations can be enhanced by carefully designing and creating them, piloting them beforehand, and revising them with an eye toward increasing their reliability before they are ever used in the main study. In cases, where humans will be rating or coding data, reliability may be enhanced by giving the raters/coders clear guidelines, carefully training them, and periodically retraining them (especially if the ratings will be done over a long period of time).

The reliability of a study's measurements and observations can be checked in cases were test items or Likert-item questionnaires are involved, either by calculating test-retest reliability (i.e., examining the degree of correlation between the scores produced by two administrations of the same test or questionnaire), parallel forms reliability (i.e., examining the degree of correlation between the scores

produced by two forms of the same test or questionnaire), or more easily, by calculating internal consistency reliability estimates (e.g., Cronbach alpha, K-R20, etc.) as appropriate. Alternatively, in cases where the measurements or observations are being assigned by raters, interrater reliability can be used (typically by examining the degree of correlation between ratings assigned by pairs of raters), and when the measures or observations are being coded by human coders, intercoder agreement will be used (typically, by calculating the percent of codings that agree between two coders).

However, at a macro level, reliability can also be defined as the degree to which the results of a study are consistent. This type of macro reliability can be enhanced by carefully (a) sampling, (b) thoughtfully planning and controlling the conditions under which the study is conducted, and (c) meticulously designing, piloting, and revising all measurement and observation tools. In general, then, the reliability of a study should be examined in terms of how well the results of the study are internally consistent and make sense in terms of sampling, study conditions, and instrumentation.

Validity

In quantitative research, at a micro level, validity can be defined as the degree to which a study's measurements and observations represent what they are supposed to characterize. The validity of a study's measurements and observations can be enhanced by carefully designing and creating them based on the best available language learning theories, piloting them beforehand, and revising them with an eye toward increasing their validity in terms of how accurately they are measuring what they were intended to measure.

The validity of the scores or other values obtained from any instrumentation in a study can be checked and/or defended by studying evidence and developing arguments for the content, criterion-related, construct validity of the resulting scores or other values, as well as their social consequences and values implications within the study and more broadly.

At a macro level, validity can also be defined as the degree to which the results of a study represent what the researcher thinks they represent. This type of macro validity can be enhanced by initially designing a study to maximally approximate "natural" conditions; by carefully prearranging and controlling study conditions; and by guarding against effects like the Hawthorne effect, halo effect, subject expectancy effect, researcher expectancy effect, practice effect, and reactivity effect (see Brown, 1988, or many other sources).

Replicability

Replicability can be defined as the degree to which a study supplies sufficient information for the reader to verify the results by replicating or repeating the study. The replicability of a study can be enhanced by writing a clear and complete research report in the style of a recipe that tells readers about: the participants (including who they were and how they were selected), the materials (including what measurements and observations were used in the study and why they were reliable and valid for that purpose), the procedures (including all of the steps in how the study was conducted), and the analyses (including how the variables were defined and arranged, as well as all analyses that were performed to address the research questions). Indeed, the study should be so clearly described that a reader could in fact repeat the study if they were so inclined. One way to check this is to ask a colleague to read the report and give you feedback with the notion of replicability (as described here) in mind.

Generalizability

Generalizability can be defined as the degree to which the results of a study can be generalized, or are meaningful, beyond the sample in a study to the population that the sample represents. Unfortunately, it

is often very difficult to define a general population in second language studies. For example, in an ESL study, can we ever say that a sample of students selected from the English Language Institute at the University of Hawai'i at Mānoa (UHM) is representative of all ESL students studying in the US? Or even all ESL university students studying in the Hawai'i? Can we say that this predominantly Asian sample of international students is the same or even similar to ESL students studying at a US East Coast university, where students might tend to be predominantly European and Middle Eastern? I think you can see the problem.

However, there is no reason to lose hope because the generalizability of a study can be enhanced in at least four ways:

- Narrowly define the population you are trying to sample from. For example, don't even pretend that you are trying to generalize to all ESL students in US universities (or even to all EFL students in Japanese universities). Instead, define the population narrowly as in the population of all students in the ELI at UHM. Then and only then will it be reasonable to say that a sample selected randomly or in a stratified manner represents that population of students in the ELI at UHM.
- Choose participants with random or stratified selection into the study and then into whatever groups you may want to compare (e.g., treatment and control groups). Those strategies will definitely help to improve the representativeness of the sample(s) and thus the generalizability of the study (see Brown, 2006).
- Control for self-selection and mortality of participants (a) by avoiding the use of volunteers whenever possible (i.e., self-selection) and (b) by minimizing as much as possible all attrition (i.e., participants dropping out of the study, also known as, mortality) by keeping the study short in duration and by encouraging participants to stay in the study. The reasoning here is that people who volunteer tend to be a certain type of gung-ho student not representative of the entire population, and similarly, people who leave a study or drop out may also be a certain type of person who by leaving will make the remaining participants less representative.
- Use the qualitative concept of transferability described in Brown (2005), which was described as follows: "Transferability can be enhanced by providing what is often referred to as thick description (i.e., giving enough detail so the readers can decide for themselves if the results are transferable to their own contexts)" (p. 32). What I am saying is that providing readers with very clear information about who the participants were and how they were selected will help those readers determine for themselves how much the results can be generalized, or better yet, how much the results may apply to their own teaching/research situations.

Conclusion

In direct answer to your original question, the characteristics of sound quantitative research are generally considered to be: reliability, validity, replicability, and generalizability.

These are of course ideals that researchers should strive for and of course may be enhanced, or defended in a variety of different ways depending on the type of study, the research questions involved, the nature of the variables, the choices of statistical analysis techniques, and so forth. Because these characteristics are ideals, they can also serve as standards against which you as a reader can judge the quality of quantitative research that you encounter in our ever growing literature. And of course, remember to apply these same standards just as critically to any research that you yourself may produce.

Those readers who find quantitative research methods intriguing may find it useful to read books like Baayen (2008), Brown (1988, 2001), Brown and Coombe (2015), Brown and Rodgers (2002), Butler (1985), Dörnyei (2003, 2007), Hatch and Lazaraton (1991), Porte (2010), Rietveld and van Hout (1993),

and Scholfield (1995); and, those interested in moving beyond the basic level should consider reading Plonsky (2015) and perhaps even Tabachnick and Fidell (2012).

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, UK: Cambridge University.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* London: Cambridge University.

Brown, J. D. (1992). What is research? *TESOL Matters*, *2*(5), 10.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University.

Brown, J. D. (2004). Research methods for applied linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 476-500). Oxford: Blackwell.

Brown, J. D. (2005). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound qualitative research. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(2), 31-33. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_22.htm

Brown, J. D. (2006). Statistics Corner. Questions and answers about language testing statistics: Generalizability from second language research samples. *Shiken: JALT Testing & Evaluation SIG Newsletter, 10*(2), 24-27. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_24.htm

Brown, J. D. (forthcoming in 2016). Characteristics of mixed methods research. *Shiken Research Bulletin, 20*(1).

Brown, J. D., & Coombe, C. (Eds.) (2015). *The Cambridge guide to research in language teaching and learning*. Cambridge: Cambridge University.

Brown, J. D., & Rodgers, T. (2002). *Doing second language research.* Oxford: Oxford University.

Butler, C. (1985). *Statistics in linguistics*. Oxford: Blackwell.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Rowley, MA: Newbury House.

Plonsky, L. (Ed.) (2015). *Advancing quantitative methods in second language* research. New York: Routledge.

Porte, G. K. (2010). *Appraising research in second language learning: A practical guide to critical analysis of quantitative research.* Amsterdam: Benjamins.

Rietveld, T., & van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour*. Berlin: Mouton de Gruyter.

Scholfield, P. (1995). *Quantifying language: A researcher's guide to gathering language data and reducing it to figures.* Clevedon, UK: Multilingual Matters.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). New York: Pearson.

Where to Submit Questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu

JD Brown
Department of Second Language Studies University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822, USA

Call for Papers

Shiken is seeking submissions for publication in the June 2016 issue. Submissions received by 1 April, 2016 will be considered, although earlier submission is strongly encouraged to allow time for review and revision. *Shiken* aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or .docx format) using 12 point Times New Roman font, although plain text files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings “Tables” and “Figures”. Tables and figures should be numbered and titled following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.* Within the body of the text, indicate approximately where each table or figure should appear by typing “Insert Table x” or “Insert Figure x” centered on a new line, with “x” replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line space from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading “References”. Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.*