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.