An analysis of the Oxford Placement Test and the Michigan English Placement Test as L2 proficiency tests

ABE, Mariko / 酒井, 英樹 / WISTNER, Brian / 阿部, 真理子 / ウィスナー, ブライアン / SAKAI, Hideki

(出版者 / Publisher)
法政大学文学部
(雑誌名 / Journal or Publication Title)
Bulletin of Faculty of Letters, Hosei University / 法政大学文学部紀要
(巻 / Volume)
58
(開始ページ / Start Page)
33
(終了ページ / End Page)
44
(発行年 / Year)
2009-03-10
(URL)
https://doi.org/10.15002/00004453

An Analysis of the Oxford Placement Test and the Michigan English Placement Test as L2 Proficiency Tests

Brian Wistner, SAKAI Hideki and ABE Mariko

Abstract

The purpose of this study was to analyze two second language (L2) English placement tests, the Oxford Placement Test and the Michigan English Placement Test, to evaluate the degree to which the tests function as L2 proficiency tests. Descriptive statistics, normality tests, and correlation coefficients were calculated for the two tests. Factor analytic methods and Rasch scaling were also applied to the data set. The results suggest that both tests can function as proficiency tests for Japanese learners of English. However, the distribution of the Michigan English Placement Test scores deviated from normality, and the reliability estimates of the scores on some of the subsections of that test were low. The implications of the results are that researchers need to choose a testing instrument that measures the aspect of proficiency that is related to a particular study, and that subsection scores may not be reliable enough to use as a unidimensional scale.

1. Introduction

Language tests vary in the purpose, functions, and characteristics of tests. The categories of language tests, for example, involve the following types: proficiency, placement, achievement, and diagnostic tests (Alderson, Clapham, & Wall, 1995; Brown, 2005) and progress tests (Alderson, Clapham, & Wall, 1995). This study focuses on one type of language test, proficiency tests.

Proficiency tests are norm-referenced and are intended to “measure global language abilities” (Brown, 2005, p. 2). One characteristic of a proficiency test, as a norm-referenced test, is that it should produce “scores which fall into a normal distribution” (p. 5), which allows relative interpretations of the test scores in terms of “how each student’s performance relates to the performances of all other students” (p. 4). A second characteristic is its test structure: The test “is relatively long and contains a wide variety of question content types” (p. 5), and usually consists of “a few subtests on rather general language skills like reading comprehension, listening comprehension, grammar, writing, and so on” (p. 5). Further, a third characteristic of the test is that “the test must provide scores that form a wide distribution so that interpretations of the differences among students will be as fair as possible” (p. 8). In other words, a proficiency test tends to test overall general language proficiency.

In the field of L2 learning and teaching, proficiency tests are often utilized to measure participants’ L2 proficiency and to divide the participants into several proficiency groups. Thomas (1994), through the analysis of a corpus of L2 acquisition studies, identified four common means of assessing L2 proficiency: impressionistic judgment, institutional status, in-house assessment instrument, and standardized

tests. Specifically, L2 standardized tests were the second most frequent means of assessment, accounting for 22.3 percent of the total corpus. The relevance of the reported usage of standardized tests becomes clear when the importance of the accurate measurement of L2 proficiency is considered–L2 proficiency is often a key variable in L2 acquisition studies. There is a need for researchers to measure L2 ability both accurately and precisely. Most notably, the comparability of the results of studies which have used L2 proficiency as a variable becomes uninterpretable if the reliability and validity of scores from the employed instruments are not reported or considered in each study.

Thomas (1994) pointed out that "There is also the important issue of what standardized tests measure, and whether what they measure is of interest in a given experimental context" (p. 326). The TOEFL, which Thomas found to be used most frequently, does not provide detailed information of the test items, so researchers have difficulty in judging whether what the test measures is relevant to their studies. Moreover, since only overall scores of the test are available, researchers can hardly assess the reliability of the test for the participants in their studies. On the other hand, the questions of the Michigan English Placement Test (MEPT, Corrigan, Dobson, Kellman, Spaan, & Tyma, 1993), which Thomas (1994) found to be the second most commonly used L2 proficiency measure in applied linguistics research, are accessible to researchers. The test may be preferable to other assessment measures in that researchers can evaluate what the test measures and whether the test is relevant to their studies, as Thomas (1994) pointed out.

The purpose of this study is to examine whether two tests, the MEPT and the Oxford Placement Test 2 (OPT, Allen, 1992) are appropriate for Japanese university students. The following research question was posited: Are the MEPT and the OPT suitable for Japanese students as proficiency tests? More specifically, the current study will answer the following questions:

(1) Are the scores of the MEPT and the OPT normally distributed?

(2) Are the reliability coefficients of the MEPT and the OPT sufficiently high?

(3) What do the MEPT and OPT measure?

(4) Do the MEPT and the OPT precisely measure a wide range of proficiency levels?

2. Method

2.1. Participants

The participants for this study were 132 university students from required English courses at two universities (University T and University C) in Japan, who took both the MEPT and the OPT. The participants from University T belonged to the Department of Economics (n = 89): Twenty-two participants were 1st year students (17 males and 5 females); 66 were 2nd year students (46 males and 20 females); and one 3rd year female student. Thus, most of the participants were male students (70.8%), and a majority of the participants were 2nd year students (74.2%). On the other hand, the participants from University C were all 1st year students majoring in engineering (n = 43): Forty participants were male students (93.0%), and three were female students (7.0%).

2.2. Michigan English Placement Test

The MEPT consists of listening, grammar, vocabulary, and reading sections, which contain 20, 30, 30, and 20 items respectively. The listening section includes 20 items. There are two types of multiple-choice listening questions with three options printed in the test booklet. First, test-takers hear short questions and are asked to choose the appropriate responses to the questions. Buck (2001) categorized this type of listening test as *response evaluation*. Second, they listen to short sentences and are asked to select the best options that express the closest meaning. This type of test is categorized as *paraphrase recognition* (Buck, 2001). Both types of questions require test-takers to understand the literal meaning of the recorded English sentences and written options. All of the item stems are short and comprised of basic vocabulary items. The options consist of three to six words; some of them are not even sentences but phrases.

The grammar section contains 30 items with four options. Test-takers are asked to read the item stem which contains one blank and to fill in the blank with one of the four options. The test type is a multiple-choice task (Purpura, 2004). The questions cover a wide rage of grammatical structures including choosing an appropriate pronoun form, verb form, or word order, and identifying the appropriate use of prepositions and prepositional phrases.

The vocabulary section consists of 30 items. The format for the vocabulary section is the same as that of the grammar section. Test-takers are asked to read the stem with one blank and to choose the word from the four options that best completes the sentence. Unlike the grammar section in which the item stems are based on a conversation between two people, the vocabulary test items are based on one or two short sentences.

The reading section includes 20 items. Test-takers read an item prompt, which varies in complexity and length across items, and includes one question about the information in the sentence. They are asked to choose an appropriate response out of four written alternatives. The average length of the items in the reading section is approximately 20 words.

2.3. Oxford Placement Test

The OPT consists of listening and grammar sections. The listening section consists of 100 items. It takes approximately ten minutes to complete the listening test. Test-takers are asked to choose the correct word which they hear in short sentences from two choices. Buck (2001) called this type of test a *phonemic discrimination task* in which the test-takers' task is to distinguish two words which differ by one phoneme.

The grammar section consists of 100 items. Fifty minutes are allotted for completion. Test-takers are asked to read the stem with a blank and to choose one of the three options for the blank. Like the MEPT, the test type is a multiple-choice task (Purpura, 2004).

2.4. Analysis

Descriptive statistics were calculated to examine the distribution of scores on the two tests. Specifically, the mean, standard deviation, skewness, kurtosis, and the standard error of skewness and kurtosis were investigated to determine the extent to which the scores were normally distributed. These

statistics were calculated with and without outliers to inspect the influence of extreme scores on the distributions. Cronbach's alpha was calculated for each subsection of the two tests. Additionally, Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were calculated to check whether any deviations from normality were statistically significant. Pearson correlation coefficients were computed to investigate the relationships among the subsections of the two tests and between the total scores. A factor analysis was carried out to reveal the extent to which the construct validity of the subsections of the tests was aligned with the purported underlying factors. These statistics were calculated in SPSS 12.0. Finally, the test scores were subjected to a Rasch analysis to examine the extent to which the range of item difficulty estimates matched the distribution of test-taker ability estimates. Rasch analyses were calculated using Winsteps 3.66.0.

3. Results

Table 1 shows the descriptive statistics for the MEPT and the OPT. The means for the two tests were similar in that both were near 60 percent. The means of the subsections were also fairly well centered when viewed individually. The variance of the MEPT was a little larger than that of the OPT. Three participants were identified as univariate outliers with z-scores in excess of +/- 3.29; one score was from the MEPT while the other two were from the OPT. After removing the three outlying scores from the data set, descriptive statistics were calculated (see Table 2) and the normality of both distributions was checked. The distribution of MEPT total scores exhibited significant skewness ($p < .05$), but the kurtosis was within normality limits. The OPT total scores did not exhibit significant skewness or kurtosis. While the skewness statistics cause concern for the distribution of MEPT scores, for large samples slight deviations from normality will become statistically significant; therefore, histograms were inspected to verify

Table 1 ***Descriptive Statistics for Each Section of the MEPT and the OPT (N = 132)***

	M	*SD*	Skewness	*SES*	Kurtosis	*SEK*	*α*
MEPT_total	58.14	8.87	-0.75	0.21	0.89	0.42	.753
listening	9.26	2.38	0.31	0.21	0.12	0.42	.236
grammar	20.70	4.00	-0.59	0.21	0.62	0.42	.677
vocabulary	19.18	3.66	-0.81	0.21	0.58	0.42	.598
reading	9.00	2.78	0.21	0.21	-0.14	0.42	.451
OPT_total	125.45	13.75	-0.73	0.21	1.02	0.42	.809
grammar	55.58	9.55	-0.61	0.21	0.16	0.42	.789
listening	69.88	6.85	-0.10	0.21	1.24	0.42	.664

Table 2 ***Descriptive Statistics for the MEPT and the OPT With Extreme Scores Excluded (N = 129)***

	M	*SD*	Skewness	*SES*	Kurtosis	*SEK*	*α*
MEPT_total	58.53	8.36	-0.52	0.21	0.04	0.42	.724
OPT_total	126.40	12.34	-0.31	0.21	-0.11	0.42	.786

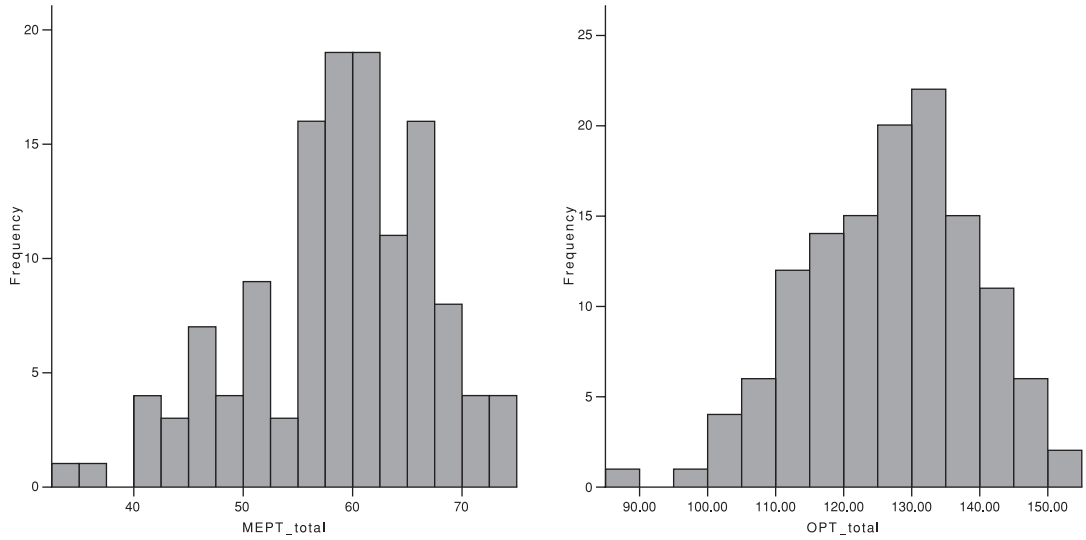


Figure 1. **The distributions of the MEPT and the OPT, N = 129.**

Table 3 ***Correlation Matrix for Test Scores and Subsections (N = 132)***

	1	1a	1b	1c	1d	2	2a	2b
1. MEPT_total		.427	.794	.782	.651	.580	.624	.294
a. MEPT_listening	.000		.085	.159	.175	.266	.247	.191
b. MEPT_grammar	.000	.335		.506	.356	.505	.559	.233
c. MEPT_voc	.000	.069	.000		.315	.408	.473	.159
d. MEPT_reading	.000	.044	.000	.000		.358	.352	.228
2. OPT_total	.000	.002	.000	.000	.000		.889	.769
a. OPT_grammar	.000	.004	.000	.000	.000	.000		.390
b. OPT_listening	.001	.028	.007	.069	.009	.000	.000	

Note. The figures to the right of the diagonal indicate correlation coefficients; the figures to the left of the diagonal indicate *p* values.

Table 4 ***Proficiency Levels According to the Given Criteria of the MEPT and OPT***

	A	B	C	D	E	Total
1	0	2	0	0	0	2
2	1	18	13	0	0	32
3	1	25	28	3	0	57
4	0	6	18	11	0	35
5	0	0	2	3	1	6
Total	2	51	61	17	1	132

Note. MEPT proficiency levels: A = advanced (low), B = intermediate, C = intermediate (low), D = beginner (high), E = beginner; OPT proficiency levels: 1 = proficient_advanced user, 2 = upper intermediate_competent user, 3 = lower intermediate_modest user, 4 = elementary_limited user, 5 = basic_extremely limited user.

any problematic deviations. Histograms for both distributions revealed no noticeable deviations from normality (see Figure 1). However, visual inspections are subjective and individual skewness and kurtosis statistics only reveal a partial picture of the distribution; that is, those statics each address only one aspect of a distribution of scores and fail to provide a test of the distribution as a whole. Therefore, normality tests were computed. For the MEPT scores, the Kolmogorow-Smirnov test, $D(129) = .10$, $p < .05$, and the Shapiro-Wilk test, $D(129) = .06$, $p < .02$, confirmed that the distribution significantly deviated from normality. For the OPT scores, the Kolmogorow-Smirnov test, $D(129) = .06$, $p = .20$, and the Shapiro-Wilk test, $D(129) = .99$, $p = .40$, confirmed that the distribution did not significantly deviate from normality. Thus, while these normality tests must be interpreted with caution do to sensitivity to sample size, any application of the MEPT scores in parametric statistical analyses requires further justification or transformation of the distribution so that it exhibits acceptable normality.

The reliability estimates for the total test scores of the MEPT and OPT were acceptable (.753 and .809, respectively, for all the participants; 724 and .786, respectively, without outliers). The reliability estimates varied greatly on the subsections of the two tests with the MEPT listening section having low reliability (.236).

The correlation between the total scores on the MEPT and the OPT ($r = .580$) was statistically significant, $p < .000$ (see Table 3); however, the coefficient of determination, obtained by squaring the coefficient, indicated that the two tests overlap 33.64%, which may lead to discrepancies in proficiency level decisions (see Table 4). Furthermore, as these two tests both purport to measure the same construct, however vaguely defined, the results suggest that the tests may measure different aspects or levels of the target construct. The correlations, however, do support the claim that the subsections of the MEPT measure different aspects of language ability.

Table 4 shows the proficiency levels that would be assigned using the criteria given in the test manuals. The proficiency judgments varied according to which proficiency guidelines were used. Specifically, the scores for the intermediate levels showed the most variation. The higher number of intermediate level learners as measured by the MEPT in relation to the OPT could be affected by differences in test difficulty. This aspect is explored in the results of the Rasch analyses.

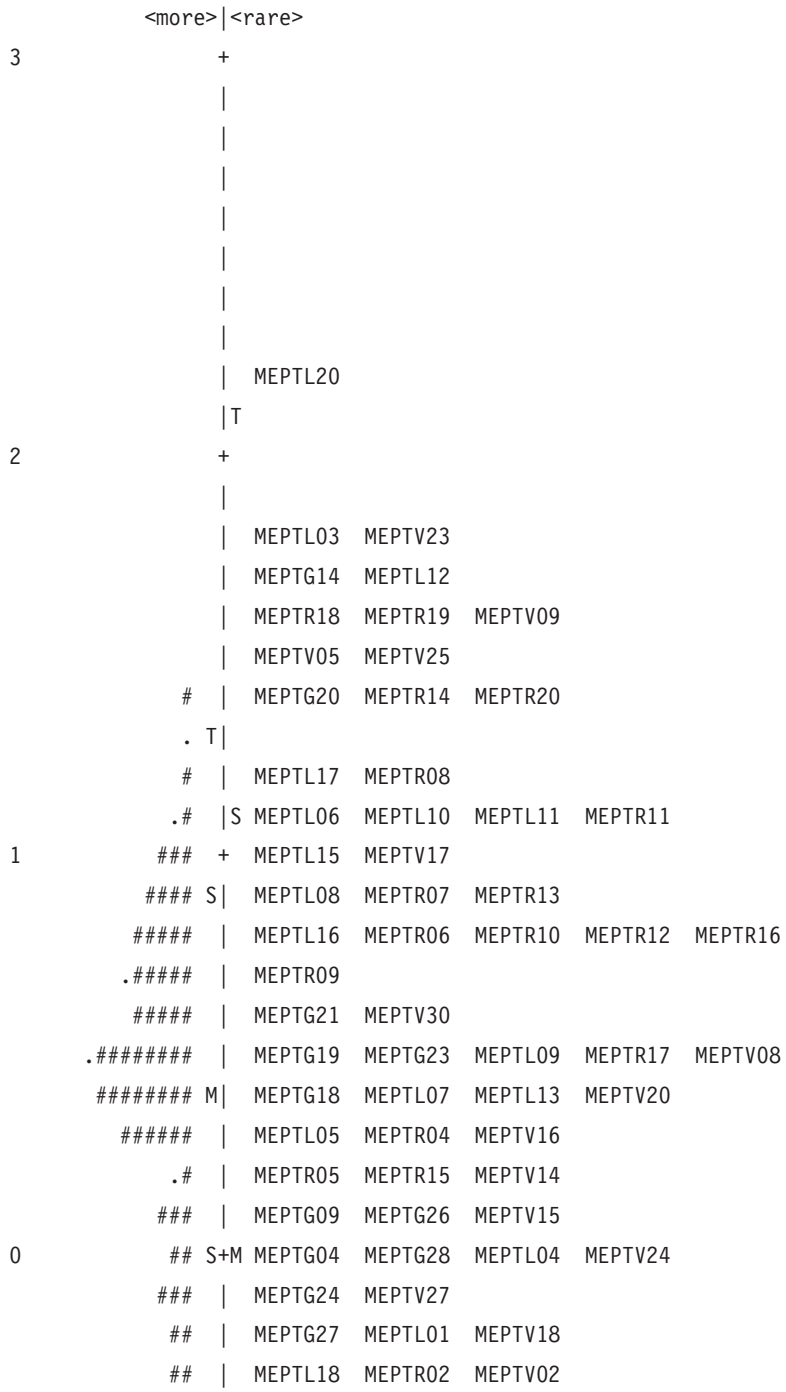
Factorial analysis yielded a two-factor solution: Factor 1 relates to grammatical and lexical knowledge, and Factor 2 relates to listening ability (see Table 5). A factor loading of .40 was used as a lower cutoff point. The reading section of the MEPT did not load on either of the two factors. The results indi-

Table 5 ***Factor Analysis Pattern Matrix***

	Factor 1	Factor 2
MEPT_listening	-0.030	0.409
MEPT_grammar	0.871	-0.116
MEPT_vocabulary	0.634	0.010
MEPT_reading	0.341	0.213
OPT_grammar	0.515	0.390
OPT_listening	0.047	0.506

Note. Principal Axis, Direct Oblimin Rotation; $N = 132$.

cate that the subsections of the MEPT measure different aspects of language ability (at least three aspects: grammatical and lexical knowledge, listening ability, and reading ability) than the OPT (two aspects: grammatical and lexical knowledge and listening ability).



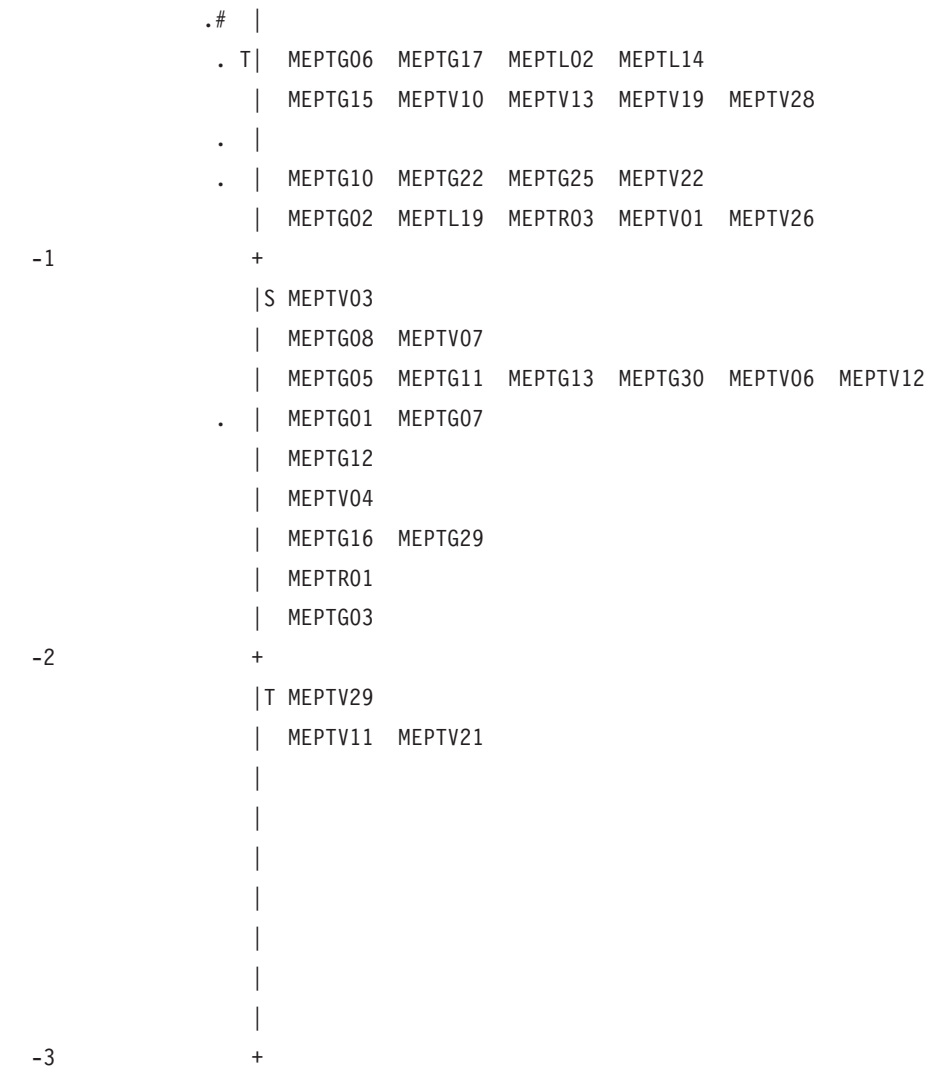
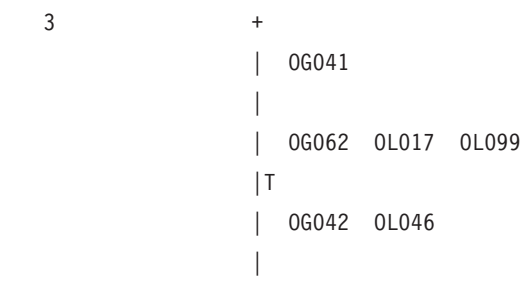


Figure 2. **Item/Person Map for the MEPT.**



```
                 |  OG008  OG079  OL047
 2               +  OL068
                 |  OG096  OL015  OL076
                 |  OG074  OG082  OL028  OL057
              .  |  OG059  OG060  OG086  OL034
              . T|  OG007
             .#  |  OG032  OG033  OG043  OG069  OG093  OL006  OL089
            ###  |S OG039  OG061  OG088  OG094  OG100  OL092
         ####### S|  OG004  OG009  OG046  OG063  OG083
 1      .#######  +  OG015  OG065  OG084  OG090  OL023
        ########  |  OG017  OG026  OG029  OG036  OG040  OG058  OG073  OG077
                  |  OG087  OG097  OL020  OL096
      ########## M|  OG071  OL009  OL053  OL059
          ######  |  OG022  OG045  OG047  OG056  OG064  OG070  OG078  OG098
                  |  OL007  OL022
         .######  |  OG018  OG023  OG034  OG052  OG072  OG085  OG089  OG099
                  |  OL014  OL026  OL058  OL079  OL090  OL094  OL095
        .####### S|  OG002  OG025  OG044  OG053  OG091  OL031
           .###  |  OG014  OG049  OG050  OG081  OG095  OL033  OL061  OL072
                  |  OL073
             ##  |  OG030  OL027  OL051  OL065  OL100
 0              T+M OG003
              #  |  OG038  OG048  OG066  OL025  OL043  OL075  OL081  OL087
              .  |  OG080  OL002  OL018  OL040  OL098
                 |  OG092
              .  |  OG001  OG005  OG019  OG020  OG024  OG035  OG055  OG057
                 |  OG068  OL011  OL048  OL050  OL070  OL084
              .  |  OG027  OL013  OL032  OL036  OL044  OL049  OL093  OL097
                 |  OG075  OL038  OL067  OL077
                 |  OG010  OG054  OL080  OL091
-1               +  OL003  OL037  OL064  OL085
                 |  OG006  OG016  OG021  OG028  OL001  OL052  OL062  OL069
                 |  OL082  OL083
                 |S OG051  OG067  OL045
                 |  OG037  OL019  OL042
                 |  OG011  OG013  OL004  OL005  OL041  OL074  OL078
                 |
                 |  OG076  OL021  OL035  OL056  OL066
                 |  OG012  OG031  OL024  OL055  OL086
-2               +
```

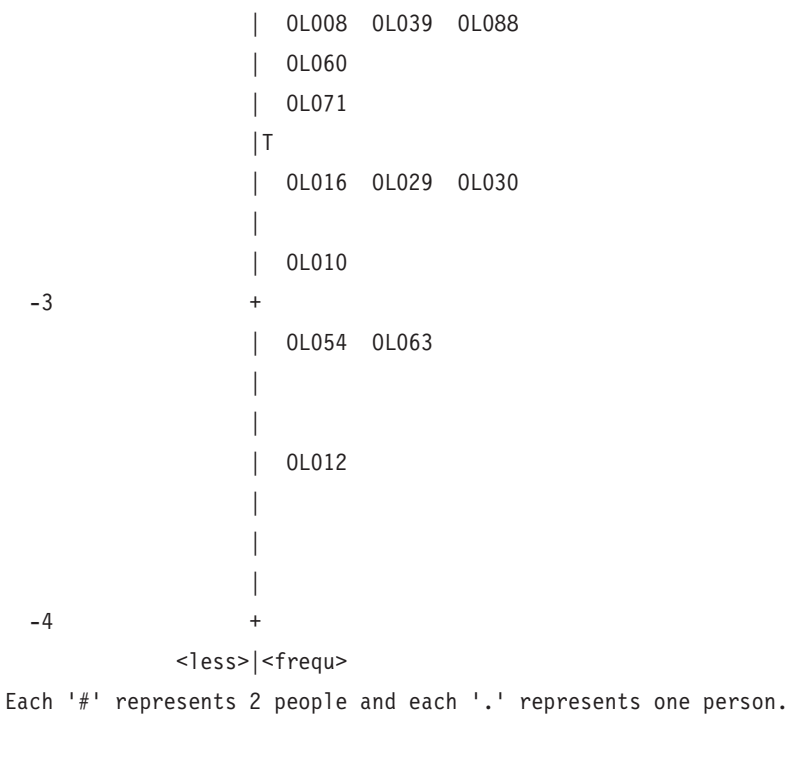


***Figure 3*. Item/Person Map for the OPT.**

The results of the Rasch analyses indicated that both tests cover a wide range of proficiency levels. The range of MEPT item difficulty estimates spanned 4. 39 logits (max: 2.22; min: -2.17). The OPT measured a larger range of proficiency with item difficulty estimates spanning 6.43 logits (max: 2.90; min: -3.53). The average error associated with the item estimates on the MEPT and OPT was acceptable at .21 and .23, respectively.

The OPT was slightly easier than the MEPT for the sample. The average person ability estimate for the OPT was .72 while the MEPT average was .42–in the case of Rasch ability estimates, higher figures represent higher levels of proficiency as measured by a certain instrument. The average error associated with the ability estimates was low: .23 for the MEPT and .17 for the OPT.

Figure 2 shows a graphical representation of the distribution of people and items. On the left side of the figure, each "#" represents the location of two participants. The scale runs from -4 logits up to 3 logits. On the right side of the figure is the distribution of test items. The items are label based on the subsection of the test and the item number. Item/person maps express statistical results in an easily comparable visual distribution similar to a histogram.

Figure 3 shows the results of the OPT. The most difficult item, OG041, is located just under the 3 logit marker. Comparing the statistics reported and the item/person map reveals that this item is estimated to be 2.90 logits, which is visually where it is located.

4. Discussion

The discussion of the results is structured around the four research questions. After situating the results, the implications of the results are discussed.

The first research question asked to what extent the distribution of scores on the MEPT and OPT are normally distributed. The OPT scores were normally distributed, thus meeting the assumption of normality that underlies most parametric statistical tests. The MEPT scores, however, exhibited statistically significant skew. This deviation would be problematic when using the scores for parametric statistical tests. If the test scores were only used for level placement, other statistical indices would be more relevant for determining the impact of using scores from a skewed distribution.

The second research question was related to the reliability coefficients estimated for the test scores. The overall test scores for the MEPT and OPT were sufficiently reliable for use in statistical tests. However, oftentimes researchers use a subsection of a test as a separate scale. In these cases, the reliability estimates for each scale need to be calculated and reported. These estimates are of specific interest when measuring a unidimensional aspect of proficiency for which a total test score may not be appropriate. In these cases, the reliability of the subsections, especially for the listening and reading sections of the MEPT, are concerning and the current estimates do not warrant use of the listening or reading sections of the MEPT as separate scales.

The characteristics of the items on the MEPT listening section could account for influencing the observed reliability. The learners that participated in Niwa, Aoi, and Yamada (2001) found the MEPT listening items to be difficult to understand due to the fast rate of speech and lack of repetition. Each item is played only once and there is little semantic repetition in the item stems.

Regarding the target constructs that the tests purport to measure, the results of a factor analysis imply that the MEPT measures a wider range of L2 English skills than the OPT. More specifically, the MEPT includes a subsection which targets L2 reading; scores from this subsection did not load on either of the two factors identified for the other subsections. Therefore, scholars need to thoroughly consider research design and the definition of L2 proficiency that informs each individual study and choose an appropriate measure of L2 proficiency.

The fourth research question sought to examine the range of person ability estimates and item difficulty estimates along with the precision with which those figures can be estimated. The item coverage for both tests was beyond adequate for the sample. The range of item difficulties exceeded that of person ability estimates, and no large gaps were observable in the distribution of item difficulty estimates. Furthermore, the error associated with the estimates was low and within expected ranges (for discussion, see Wright, 1977) at all points on the scaled variables.

Overall, the results of the current study support the use of the MEPT and the OPT with Japanese university student as tests of L2 English proficiency. The scores for the tests were reasonably reliable and there were no significant gaps in the item coverage compared to the person ability estimates. However, if a subsection of a test is used in a study as a measure of L2 proficiency, the distribution and reliability of the scores should be thoroughly examined. One way in which the reliability of the MEPT listening section could be increased would be to combine the listening sections from the three test forms of

the MEPT–this would result in a 60 item listening test, thus a higher level of reliability could be expected due to the increased number of test items. Additionally, the different aspects of L2 proficiency that the two tests measure need to be considered–the OPT measures mainly two aspects of L2 proficiency while the MEPT measures three aspects. Thus, it is important to choose a measure of L2 proficiency that is appropriate for a certain research design.

Even though the listening sections of the two tests loaded on the same factor, low coefficients indicate that the two listening sections might tap different aspects of a L2 listening proficiency construct and that the two sections do not account for much of the variance associated with the resultant proficiency scores. One possible explanation for the weak relationship observed between the two listening sections is that the MEPT uses American English and the OPT uses British English. The extent to which the students were familiar with or had been exposed to differing varieties of English was not controlled for in the present study. Thus, this difference could be one source involved in suppressing the correlation coefficient between the two listening subsections. Furthermore, the two sections measure different levels of listening processes: the OPT tests phonetic discrimination in word recognition whereas the MEPT tests the understanding of the literal meanings of sentences and the grammatically or pragmatically appropriate response to the item stem. Therefore, the type of listening proficiency that is theorized to be related to the variables under observation should be consider when selecting a testing instrument.

5. Conclusion

The present study examined the extent to which the MEPT and OPT can be used as L2 proficiency tests with Japanese learners of English. Both tests functioned well as proficiency tests, but problematic aspects were identified, namely the normality of the MEPT distribution and the reliability of the subsections. Further research is needed to examine the implications of using subsection scores as L2 proficiency measures in various research designs. Investigating the warrants and threats to the validity of different operationalizations of L2 proficiency will result in a better understanding of the ramifications of employing narrow and wide definitions of L2 proficiency in empirical research.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Allen, D. (1992). *Oxford placement test 2* (New edition). Oxford University Press.

Brown, J. D. (2005). *Testing in language programs*. New York, NY: McGraw-Hill.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Corrigan, A., Dobson, B., Kellman, E., Spaan, M., & Tyma, S. (1993). *English placement test*. The Testing and Certification Division, English Language Institute, University of Michigan.

Niwa, Y., Aoi, K., Yamada, S. (2001). The comparison between Michigan placement test lisening part and eiken pre-2nd grade listening part: The analysis of difference factors. *Chubu University Journal of International Relations, 26*, 67-86.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.

Thomas, M. (1994). Assessment of L2 proficiency in second language research. *Language Learning, 44*, 307-337.

Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116