Rasch analyses of English language placement tests

Brian Wistner
Hosei University
Hideki Sakai
Shinshu University

Reference data:
Wistner, B., & Sakai, H. (2008). Rasch analyses of English language placement tests.
In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*. Tokyo: JALT.

The purpose of this study was to apply the Rasch model to investigate the reliability of scores on English language placement tests and the validity of basing placement decisions on those scores. Previous studies have suggested that placement decisions can be carried out more reliably by creating an in-house placement test or by scoring a subset of the test items which were functioning well. In this study, we examined an in-house placement test and the Michigan English Placement Test (MEPT). The results of the analysis of the in-house placement test indicated that the error associated with the estimated ability measures around the cut score was too large to reliably divide the sample into two groups. The results of the analysis of the listening section of the MEPT revealed that although the test items exhibited good fit with the Rasch model, the items were not matched well to the sample.

本研究の目的は、ラッシュ・モデリングを用いて英語プレイスメントテストの得点の信頼性と、その得点に基づいてプレイスメントに関する判断をすることの妥当性を検討することであった。先行研究によれば、より正確なプレイスメントの判断を行うために、学習者の能力に対応する困難度を持つテストを自作するか、または、テスト項目をすべて用いるのではなく学習者の能力に対応する困難度を持つテスト項目のみを使って分析するかのいずれかの方法を用いることが提案されている。本研究では、ある大学の自作のテストと、既成のミシガン・テストを分析した。自作のテストの分析結果によれば、受験者を2つのグループに分けるためにプレイスメントの基準点を平均点に設定したとき、その前後の能力推定値の誤差は大きかったことがわかった。これは、プレイスメントに関する判断の信頼性が低いことを示している。ミシガン・テストのリスニングセクションの分析結果によれば、テストの項目はラッシュ・モデルとの適合度が良好であったが、項目の困難度推定値は受験者の能力推定値に適していなかったことがわかった。

Tests serve important functions in many language programs. From course grades to program-level placement, scores on language tests often form the core of the decision-making processes. Placement tests, in particular, directly affect numerous stakeholders in a language program. A common use of placement tests is to "assess students' level of language ability so that they can be placed in appropriate course or class" (Alderson, Clapham, & Wall, 1995, p. 11). Oftentimes students are placed into a level or a class based on a single test score, but the reliability of the scores is rarely investigated. Furthermore, evidence for the validity of basing placement decisions on a certain test score is seldom gathered or examined. The purpose of this study is to examine the recommendations reported in the literature on how to improve placement tests and to apply the Rasch model to investigate the reliability and validity of using certain placement tests for placement decisions in language programs.

Placement tests perform a unique function when viewed among the range of different test types. Brown (2005) listed four different types of tests commonly used in language programs. Placement tests can be used to test incoming and continuing learners to determine their level within a language program. Achievement tests can be used to test learners' abilities, usually at the end of a course. Proficiency tests aim to measure the overall language proficiency of a learner, and diagnostic tests can be used to assess learners' control of certain language features or structures. Alderson, Clapham, and Wall (1995) mention another type of test: progress tests. Progress tests are similar to achievement tests, but they are commonly given throughout a course to measure learners' progress at different points in time.

Placement tests are often thought to be the same or similar to proficiency tests. While in some situations this may be the case, considerable differences can be found between the two test types. Norm-referenced proficiency tests by and large seek to "measure global language abilities" (Brown, 2005, p. 2). Characteristics of norm-referenced placement tests include items that attempt to assess a wide range of abilities based on criteria that are not related to a specific language program. Many standardized tests are norm-referenced. Placement tests, however, attempt to assess a narrower range of abilities to group students within a program. For instance, learners entering a language program might have scores within a certain range on a norm-referenced proficiency test, and then the language program would use a placement test to further divide that group of learners in a meaningful way so that appropriate class placement can be made. Thus, proficiency tests tend to test overall general language proficiency, and placement tests tend to focus on a smaller range of skills or knowledge that are normally related to the language program's curriculum.

When choosing or designing a placement test, a number of considerations must be deliberated. First, the difficultly of the placement test should fit the students' ability levels. Having numerous test items that are at similar difficulty levels as the students' ability levels will increase measurement precision and provide meaningful information for placement decisions. Second, the test should be accurate. It should produce reliable scores so that based on the scores teachers and administrators are able to "place students into the appropriate levels with little or no error" (Murray, 2002, p. 22).

A limited number of previous studies have investigated these considerations when using commercially produced proficiency tests for class placement decisions in Japanese universities. Culligan and Gorsuch (1999) used classical test theory to examine the functioning of the SLEP test for placing first-year university students into class levels. As the test items did not function well for their purposes, they recommended scoring a subset of items that exhibited well-centered item facility statistics and that discriminated at the .30 level or higher. Using a subset of the test items for placement can help to lower the standard error of measure (Sakai & Wistner, 2007), but this method of test scoring becomes problematic when too few items are found to exhibit item facility and item discrimination statistics within reasonable levels (see Brown, 2005, for a discussion of item facility and item discrimination ranges). For instance, Abe, Wistner, and Sakai (2008) found that less than half of the items on the listening section of a commercially produced placement test were functioning at expected levels when analyzed using classical test theory.

Another recommendation for test revision found in previous studies is that language programs should make in-house placement tests. Westrick (2005) concluded that creating an in-house test would be a sound solution to overcome the poor performance of a commercially produced proficiency test.

Gorsuch and Culligan (2000) sought to overcome the limitations of classical test theory by using the Rasch model to analyze a placement test. The results indicated that the Rasch ability and difficulty estimates provided useful information about the test takers and that placement decisions could be refined.

In summary, previous studies have found that the commercially produced proficiency tests that have been examined do not work well when used for placement purposes in Japanese universities. Recommendations for placement test revision have been that either an in-house placement test should be created, or a subset of items should be scored. In this paper, we report the results of a Rasch-based investigation into an in-house placement test and a commercially produced placement test, the Michigan English Placement Test (MEPT).

Research questions

Two research questions were posited for this study.

1. For the in-house test, how precise are the ability estimates of the test-takers around the cut point for placement decisions?
2. For the MEPT, how well does the listening section of the MEPT fit the students' ability levels (for placement into listening and speaking classes)?

More specifically, the second research question seeks to determine how precise the item estimates are and how well the test items fit the Rasch model.

Method

Participants and placement tests

Data for the analysis of the in-house placement test came from all the 2nd-year university students of the Faculty of Education in a Japanese national university ($N = 283$). The data were collected in the beginning of the 2006 academic

year. The in-house placement test was administered for the purpose of placing students into two levels (advanced and intermediate) for the required general English courses. The test consisted of three sections in the multiple-choice format: 10 items for listening, 10 items in cloze tests, and 10 items for grammar (see Sakai & Wistner, 2007, for a more detailed description of the test).

Form C of the MEPT was administered at the beginning of the 2005 Fall Semester at a private Japanese university (N = 149). The breakdown of the participants by school year is as follows:

1st year students	(n = 52)
2nd year students	(n = 43)
3rd year students	(n = 24)
4th year students	(n = 28)
Study abroad students	(n = 2)

The listening section the MEPT contains 20 items. Test takers listen to a statement and then choose the best response from three choices written in their test booklets. The remaining section of the MEPT consists of 30 grammar items, 30 vocabulary items, and 20 reading items.

Analyses

To answer the research questions, descriptive statistics and Rasch person ability and item difficulty estimates were calculated for scores on the in-house placement test and for scores on the MEPT. Within the dichotomous Rasch model, linear person ability and item difficulty estimates can be calculated, along with the error estimates for each person and item estimate. The person ability measures and item difficulty measures are obtained by a log-odds transformation of the raw scores into a continuous scale. The ability and difficulty measures are then reported as logits. Logits can be thought of as units of measurement based on an interval scale. Descriptive statistics and correlations were calculated in SPSS 14, and Winsteps 3.6.0 (Linacre, 2006) was used for application of the Rasch model to the data sets.

Results

The results of the in-house placement test analysis are reported first and are followed by the results of the analysis of the listening section of the MEPT.

In-house placement test

Table 1 shows the descriptive statistics for the in-house placement test. While the measures displayed fairly good distribution, the mean of the person measures were higher than 0, which is the arbitrarily set mean of the item measures. The confidence intervals are reasonably tight, with only a 0.2 logit spread around the mean. The person separation is somewhat low considering that the test scores are used for placement purposes. A person separation statistic of 2 or higher is desirable for measures which are used for grouping (Bond & Fox, 2007).

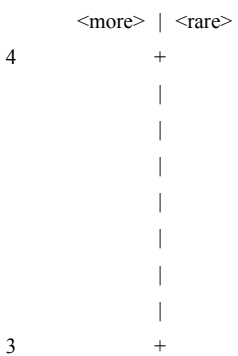
Table 1. Descriptive statistics for the in-house placement test (N = 283)

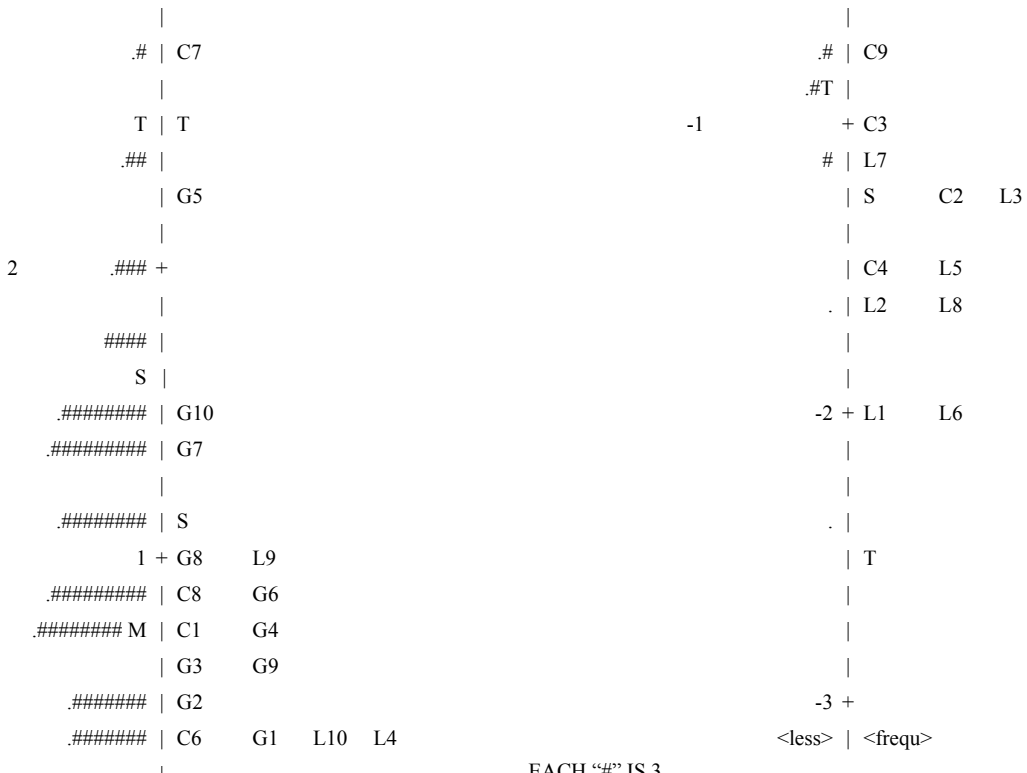
k	30
M	.80
95% CI Lower	.70
Upper	.90
SD	.85
Item Reliability	.98
Item Seperation	7.74
Person Reliability	.68
Person Separation	1.47
Skewness	-.19
SE of Skewness	.15
Kurtosis	.44
SE of Kurtosis	.29

We analyzed the three sections as one, mainly because the total scores of the three sections are treated as an indication of general English ability for the actual placement decisions. To examine the unidimensionality, we performed a principle component analysis of the residuals, examined item misfit, and calculated the correlations among the person ability measures on each section. Although three of the 30 items underfit the Rasch model, the remaining items showed good fit. The principle component analysis of residuals did not indicate multidimensionality. Additionally, the three sections were statistically significantly correlated. Based on these findings, we treated the three subsections as one.

Figure 1 shows the distribution of the difficulty estimates of the 30 items of the in-house placement test. The distribution of items is shown on the right side, and the distribution of person ability estimates is shown on the left side. The items at the top of the map are more difficult, and items at the bottom are easier. Likewise, participants at the top of the map are estimated to have more ability compared to the participants located in the lower part of the map. The mean of the item difficulty estimates is arbitrarily set at 0.

The mean ability estimate ($M = 0.80$, $SD = 0.85$) is much larger than the mean item difficulty estimate ($M = 0.00$, $SD = 1.25$). About ten test items cluster together between -2.00 logits and -1.00 logits, where few persons are found. On the other hand, a number of persons are observed above 1.0 logits; however, there are only four items (cloze #7, grammar #5, grammar #10, and grammar #7) with difficulty estimates in that range. In other words, the test was easy for this sample of students.





EACH "#" IS 3.

Figure 1. Item map for the in-house placement test

To examine the precision of the placement decisions, the error estimates of the person ability measures around the cut point were examined. The mean raw score (M = 19.1 out of a possible 30 points) was set as a cut point. The ability

estimate of those who had raw scores of 19 points was 0.73 with the error estimate being 0.43; on the other hand, the ability estimate of those who had raw scores of 18 points was 0.54 with the error estimate being 0.43. Considering the large error estimates (0.43 and 0.43) compared to the difference in the ability measures (0.73 and 0.54), the division of the students around the cut point may have been carried out by chance.

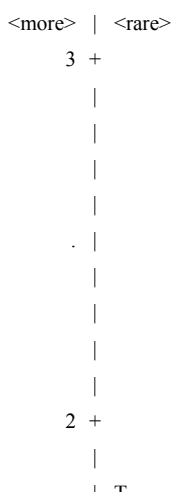
MEPT

Table 3 shows the descriptive statistics for the listening section of the MEPT. The mean ability measure is a little more than a half logit (-.58) below the mean of the item difficulty estimates. The person separation is somewhat worrisome at .75. Despite the spread of ability estimates (-2.51 to 2.50 logits), two statistically distinct groups are not observable in the data.

Table 3. Descriptive statistics for the listening section of the MEPT (N = 149)

K	20
M	-.58
95% CI Lower	-.69
Upper	-.47
SD	.68
Item Reliability	.95
Item Seperation	4.50
Person Reliability	.36
Person Separation	.75
Skewness	.49
SE of Skewness	.20
Kurtosis	2.41
SE of Kurtosis	.40

Figure 2 shows the distribution of item difficulty estimates in relation to the distribution of person ability estimates. Ten of the twenty items cluster between .26 and 1.37 logits, while only nine participants are estimated to be in that ability range. Thus, half of the listening items are not targeted well to the sample.



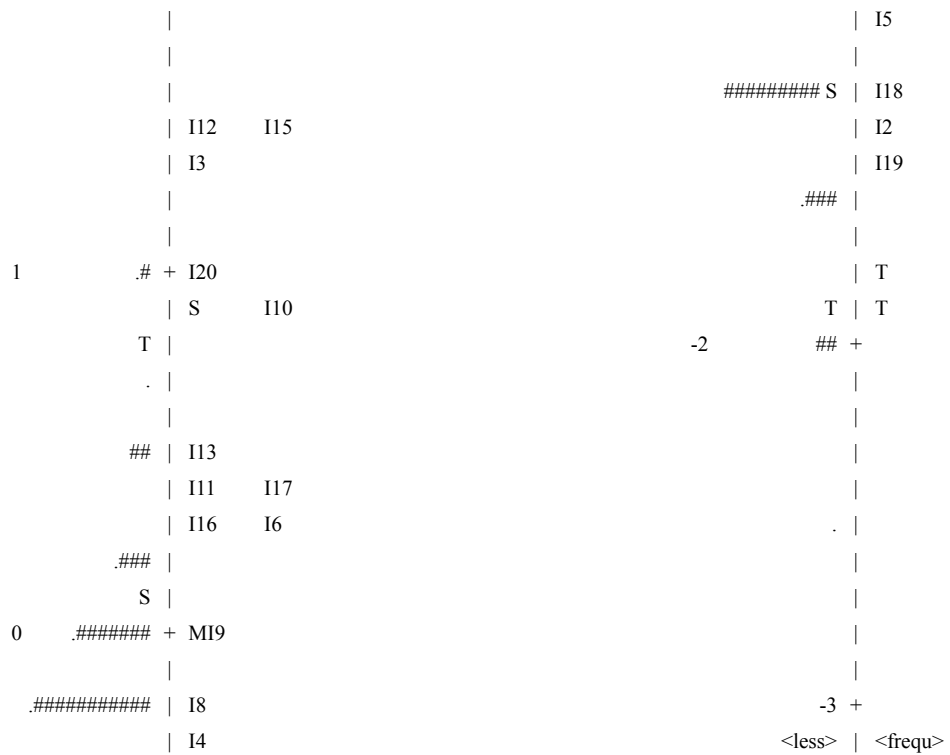


Figure 2. Item map for the listening section of the MEPT

Table 4 shows the difficulty estimates, error estimates, fit statistics, and point-biserial correlations for the 20 items of the listening section of the MEPT. Using the range of

-2.0 to 2.0 (standardized) for judging misfit, none of the items exhibited misfit with the expected values of the Rasch model. Furthermore, no items had negative point-biserial correlations. Regarding the error associated with the item estimates, all of the error estimates were below .25, which is consistent with the expected values of a sample size over 100 (Wright, 1977).

Table 4. Results of the Rasch analysis of the 20 listening items (measure order)

			Infit		Outfit		
Item	Measure	Model S.E.	MNSQ	ZSTD	MNSQ	ZSTD	PTMEA Corr.
12	1.37	.24	.92	-.4	.78	-1.0	.38
15	1.37	.24	.95	-.3	.87	-.6	.33
3	1.26	.23	1.07	.5	1.22	1.1	.11
20	.96	.21	1.06	.5	1.05	.4	.18
10	.91	.21	1.03	.3	1.10	.7	.20
13	.47	.19	.92	-.9	.83	-1.5	.42
11	.43	.19	1.05	.6	1.09	.8	.20
17	.43	.19	1.06	.8	1.09	.8	.19
16	.29	.19	.97	-.3	.98	-.2	.32
6	.26	.18	.96	-.5	.96	-.4	.33
9	.03	.18	.94	-1.0	.95	-.6	.38
8	-.16	.17	.98	-.4	.95	-.7	.33
4	-.31	.17	.99	-.3	1.00	.0	.31
14	-.57	.17	1.02	.5	1.02	.3	.26
1	-.75	.17	.98	-.3	.98	-.4	.32
7	-.75	.17	.98	-.3	.97	-.5	.32
5	-1.11	.18	1.10	1.7	1.14	1.7	.13
18	-1.30	.18	.97	-.5	.96	-.4	.33
2	-1.37	.18	1.02	.4	1.03	.3	.24
19	-1.47	.18	1.00	.0	1.00	.0	.27

Discussion

The results of the analysis of the in-house placement test showed that (a) the errors of the person ability estimates around the cut point were too large for clear division, and (b) the test items were easy for the sample. One probable reason for the large error estimates around the cut point is that there were few items around the cut point. Only three items (grammar #4, grammar #9, and grammar #3) fell in this range (see Figure 1). For revision of the test, it is necessary to add more items around the cut point to reduce the error estimates of the items at and around the cut point. Furthermore, the addition of more items with relatively high difficulty estimates will be required so that the test difficulty may fit the students' ability levels.

The results of the analysis of the listening section of the MEPT revealed that the items are functioning well. The range of the difficulty estimates spanned 2.84 logits, no items exhibited misfit with the Rasch model, and the error estimates were low for all items. However, the person separation statistic is cause for concern. At .75, only one group is observable in the data. One possible explanation is that this group of learners has similar ability levels; thus, a finer-tuned instrument is needed to produce measures that meaningfully divided the group into different levels of ability. The purpose of a placement test is to measure learners' abilities in order to place them into a course. If the test is unable to create statistically distinct groups with low error estimates around the cut scores, then the validity of basing decisions and interpretation on the placement test scores comes into question. If the listening section of the MEPT were to be used as a placement test for similar groups

of learners, more items that fit the learners' levels would be necessary to increase reliability and the Rasch person separation index. While it is possible to rewrite or add items, time might be better spent creating a placement test that is related to the curriculum and goals of the language program in which it is used.

Conclusion

The goal of this study was to apply the Rasch model to examine an in-house placement test and a commercially produced placement test in order to assess the degree to which the tests produced reliable scores which are useful for placement purposes. While the commercially produced placement test consisted of items that performed well psychometrically, the resultant measures were not meaningful enough to reliably divide the sample into distinct groups. Likewise, the in-house placement test exhibited high error estimates around the cut point, which implies that placement decisions might have been carried out by chance. The implications for program-level placement decisions are that learners are often arbitrarily placed into a class even though teachers assume that scores from placement tests are reliable enough for placement decisions. Using a general proficiency test for placement could result in a wide range of test scores, but there may not be meaningful differences between many of the scores. That is, students are placed into a level based on decisions informed by test scores which exhibit low reliability, not on test scores that exhibit high reliability and adequate Rasch person separation statistics. One obvious concern is that teachers assume that the learners are of the same or similar level, when in reality there may be large discrepancies in observed abilities. Moreover, placement decisions may adversely affect students if they are placed in a level that is too high or too low. Regardless of the chosen placement test, test items need to be examined and evaluated at regular intervals to assess item functioning, reliability, separation, and the validity of basing decisions on the test scores.

Brian Wistner teaches at Hosei University. His current research interests include instructed SLA, the roles of input and output in SLA, and task-based language teaching. <wistner@hosei.ac.jp>

Hideki Sakai teaches at Shinshu University. His current research interests include second language interactional studies, classroom SLA, and psycholinguistic aspects of listening and speaking. <sakaih@shinshu-u.ac.jp>

References

Abe, M., Wistner, B., & Sakai, H. (2008). Analyzing an English language proficiency test using classical item analysis and Rasch modeling. *The Economic Journal of Takasaki City University of Economics, 50*, 125-134.

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

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Culligan, B., & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal, 21*, 7-25.

Gorsuch, G., & Culligan, B. (2000). Using item response theory to refine placement decisions. *JALT Journal, 22*, 315-325.

Linacre, J. M. (2006). WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis (3.63.0). [Computer software]. Chicago: MESA Press.

Murray, J. (2002). Creating placement tests. *ESL Magazine, November/December*, 22-24.

Sakai, H., & Wistner, B. (2007). Classical item analysis of an in-house English placement test: Issues in appropriate item difficulty and placement precision. *JACET Chubu Journal, 5*, 13-27.

Westrick, P. (2005). Score reliability and placement testing. *JALT Journal, 27*, 71-93.

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