Classical item analysis of an in-House English placement test: issues in appropriate item difficulty and placement precision

Hideki Sakai
Shinshu University

Brian Wistner
Tokyo Junshin Women's College

要旨

本研究の目的は、古典的項目分析を用いてテスト項目の難易度の適切性とプレイスメント意思決定の正確さについて調べることである。大学2年生に対して実施されたプレイスメントテストのデータ（N= 283）を分析した。テストは、リスニング問題、クローズテスト、文法問題の各10問から成る多肢選択問題であった。古典的項目分析として、項目容易度(item facility）と項目弁別力（item discrimination）が計算された。項目容易度は、各項目の正答率であり、項目の難易度を示す指標である。一方、項目弁別力は、合計得点の上位群の正答率と下位群の正答率の差である。結果は次の通りである。まず、分析対象のプレイスメントテストは受験者にとって容易であったことが示された。次に、項目容易度と項目弁別力の点から15項目選び、修正データセットを作ったところ、元データセットの測定の標準誤差（standard error of measurement）よりも修正データセットの測定の標準誤差のほうが小さかった。このことより、修正データセットを用いた場合、プレイスメントの意思決定の信頼性が高くなることが示唆された。また、受験者を2つの集団に分けるプレイスメントの意思決定について分析した。元データセットを用いた場合と修正データセットを用いた場合でプレイスメントの判断が異なった人数は、統計的に有意でなかったが、283人中37人（13.1%）見られた。これらの結果に基づき、プレイスメントテストの分析における古典的項目分析の課題が考察された。

Keywords: placement test, classical item analysis, placement decision making
キーワード: プレイスメントテスト, 古典的項目分析, プレイスメント意思決定

Introduction

This paper reports a study of the analysis of an in-house English placement test

administered in the Faculty of Education of a Japanese national university within the framework of classical item analysis. In particular, we focus on the issue of the use of the original data set for making placement decisions.

Placement Tests

Language tests can be classified into several types according to the purpose of test: placement, achievement, proficiency, and diagnostic tests (Alderson, Clapham, & Wall, 1995; Brown, 2005) and progress tests (Alderson, Clapham, & Wall, 1995). According to Alderson, Clapham, and Wall (1995), the goal of placement tests, which is the focus of this paper, is to "assess students' level of language ability so that they can be placed in appropriate course or class" (p. 11). Placement tests are similar to proficiency tests in that both are norm-referenced tests, which are designed "to measure global language abilities" (Brown, 2005, p. 2); however, one difference may be that placement tests must assess a narrower range of abilities in order to group students efficiently within a program, whereas proficiency tests "will tend to be very, very general in character" (Brown, 2005, p. 10). Thus, Brown (2005) pointed out that the effectiveness of a placement test depends on "the degree to which that test [placement test] fits the ability levels of the students" (Brown, 2005, p. 10). In addition, Murray (2002) pointed out that placement tests should be "accurate" so that they "place students into the appropriate levels with little or no error" (p. 22).

Previous Studies on Placement Tests

So far, the degree to which a particular placement test fits the ability levels of the students and the degree to which participants are divided into appropriate levels have been examined in terms of classical item analysis (e.g., Brown, 1989; Culligan & Gorsuch, 1999; Westrick, 2005) and the Rasch model (e.g., Fujita, 2005; Fulcher, 1997; Gorsuch & Culligan, 2000). Because the present paper employed classical item analysis, we will review the studies on placement tests that used classical item analysis.

Classical item analysis primarily involves item facility analysis and item discrimination (Brown, 2005, pp. 66-76). Item facility (IF), or item difficulty, is "the percentage of students who correctly answer a given item" (Brown, 2005, p. 66). An

acceptable IF ranges from .30 to .70 (Brown, 2005, p. 75). Item discrimination (ID) is "a statistic that indicates the degree to which an item separates the students who performed well [e.g., the upper third] from those who did poorly [e.g., the lower third] on the test as a whole" (Brown, 2005, p. 68). Referring to Ebel (1979), Brown (2005) considered items with an ID of .40 and up to be "very good items," items with an ID of 30 to .39 to be "reasonably good, but possibly subject to improvement," items with an ID of .20 to .29 to be "marginal items, usually needing and being subject to improvement," and items with an ID of .19 and below to be "poor items, to be rejected or improved by revision" (Brown, 2005, p. 75).

Several studies have reported the results of classical item analyses of L2 placement tests (e.g., Brown, 1989; Culligan & Gorsuch, 1999; Westrick, 2005), while others have indirectly reported the analyses as a part of larger validation studies (e.g., Wall, Clapham, & Alderson, 1994). Brown (1989) and Wall, Clapham, and Alderson (1994) analyzed institutionally developed placement tests, whereas Culligan and Gorsuch (1999) and Westrick (2005) focused on commercially created placement tests. First, Brown[1] (1989) analyzed the scores on the reading comprehension test of 61 participants from a pool of 194 L2 students who had taken the institutional placement test of the University of Hawaii at Manoa. Although the test contained five subtests (the academic listening test, dictation, cloze, writing sample, and reading comprehension test), he analyzed only the scores of the reading section. The reading section consisted of 10 reading passages followed by a total of 60 multiple-choice questions. He compared the original data set with a revised data set containing items whose IF ranged between .30 and .70 and whose ID was .30 or above.[2] The results showed that the original data set of 60 items had a mean of 33.84, a standard deviation of 6.62, a Kuder-Richardson formula 20 reliability coefficient of .79, and a standard error of measurement of 3.52, whereas the revised data set of 35 items had a mean of 18.90, a standard deviation of 4.60, a Kuder-Richardson formula 20 reliability coefficient of .63, and the standard error of measurement of 2.79. He stated that the revised data set "is well centered (*M*) and produces a respectively wide spread of scores (*SD*)" (p. 79) and "is also reasonably reliable, especially in view of its new shorter length" (pp. 79-80).

Culligan and Gorsuch[3] (1999) examined the suitability of a commercially

produced proficiency test (the SLEP proficiency test developed by Educational Testing Service) for placement purposes. They obtained SLEP test scores from 748 students first-year students enrolled in the university and junior college divisions of one school in Japan. First, they analyzed the scores in terms of classical item analysis such as IF and ID and found that 84 items of the 150 yielded an ID value of .19 or below. Second, they compared the original data set with a new data set which contained items of high ID (.20 or over) and found that the high ID data set obtained a slightly higher reliability coefficient and a lower standard error of measure than the original data set. Based on the results of the two analyses, they suggested that the scoring of all the test items of the SLEP test should be avoided; rather, only the test items with high ID values should be scored. Thus, in general, classical item analysis seemed to address the issue of matching test difficulty and learner ability.

Lastly, Westrick (2005) examined the effectiveness of the Quick Placement Test-Pen and Paper Test (QPT-PPT) when used for placement purposes. One-hundred-sixty-one first year university students took both versions of the QPT-PPT back to back. A counter-balanced design was implemented in which one group took form one and then form two of the QPT-PPT, and the second group took form two and then form one. The results showed that the QPT-PPT test scores did not effectively distinguish high-level and low-level students. Scores were grouped tightly, offering little information for placement purposes. IF and ID values for the test items were very low. The combined group score showed that only 46 out of 120 test items had IFs between .30 and .70. The majority of the IDs for the test items were negative. Additionally, the two versions of the test had weak correlation coefficients (Group 1, r = .35; Group 2, r = .49). Considering these results, Westrick found little value in using the QPT-PPT for placement purposes. Students' scores were too tightly grouped and test items performed too poorly to offer any insights as to the students' proficiency levels. He recommended that each school produce its own in-house placement test, and that more studies of commercially-produced proficiency tests were needed.

The results of these studies point toward the need for language programs to investigate the reliability and effectiveness of their placement tests. While test scores may appear to be useful and trustworthy measures for placement purposes, they are only approximations of test-takers' true scores; thus, how test items are functioning

and the amount of error associated with the test need to be investigated. Previous studies also demonstrated effective ways of applying classical test theory to test construction and revision, offering ideas on which items to keep, which items to revise, and ways in which placement tests could be scored in order to increase reliability and reduce overall error. The purpose of this study is to apply and extend these concepts to an in-house placement test.

Research Questions

For this study, we utilize classical item analysis to answer the following research questions:

1. To what extent does the difficulty level of the placement test fit the ability levels of the test-takers?
2. To what extent are the placement decisions accurate?

Method

Participants

We analyzed the in-house placement test of the Faculty of Education of a national university in central Japan. Students of the Faculty of Education are required to take placement tests twice: at the beginning of their 1st and 2nd years in school. The placement tests are administered for the purpose of placing students into two levels (advanced and intermediate) for the required General English Courses. Of the two placement tests, we analyzed the scores on the placement test administered to 283 2nd-year university students (122 males and 161 females) in the beginning of April, 2006.

The Placement Test

This section gives a brief description of the English placement test, the scoring procedure, the placement decision making procedure, and a small segment of the data set.

The English placement test consists of three sections: 10 listening items, 10 multiple-choice cloze-type items, and 10 grammar items. In this paper, we refer to each question item by the section and the number. For example, the fifth item of the

listening section will be called *listening #5* in the text of this paper and will be indicated as *L5* in the examples and tables that follow.

In the listening section, test-takers listen to a question prompt, then a short conversation or passage, and then the same question again. Next, they read the question and the four alternative answers on the test sheet and choose the best answer to the question based on the conversation or passage. The number of the conversations and passages is 10. The following is the example question provided to the test-takers.

Test-takers hear:

How was Julie's weekend?

A: *Hey, Julie, did you have a good weekend?*

B: *It was OK.*

A: *What did you do?*

B: *Nothing much. I slept all day Saturday and watched TV on Sunday, but I really enjoyed it.*

How was Julie's weekend?

Test-takers read:

How was Julie's weekend?

A. terrific B. enjoyable C. boring D. great

The correct answer:

B

In other words, each conversation or passage has one question. Thus, the questions are independent from each other.

In the cloze section, test-takers read two passages, each of which contains five blanks. Four words given as alternatives to each blank are provided on the test-sheet. Test-takers choose the best word for the blank. The questions of C1 to C5 are given as follows:

Reading passage:

Good smiles ahead for young teeth

Older Britons are the worst in Europe when it comes to keeping their

teeth. But British youngsters (C1:) more to smile about because (C2:) teeth are among the best. Almost 80% of Britons over 65 have lost all or some (C3:) their teeth according to a World Health Organisation survey. Eating too (C4:) sugar is part of the problem. Among (C5:), 12-year olds have on average only three missing, decayed or filled teeth.

Word choices:

C1:	A. getting	B. got	C. have	D. having
C2:	A. their	B. his	C. them	D. theirs
C3:	A. from	B. of	C. among	D. between
C4:	A. much	B. lot	C. many	D. deal
C5:	A. person	B. people	C. children	D. family

The correct answers:

C1: have C2: their C3: of C4: much C5: children

Even though blanks are created in the passages, a closer look at each item suggests that each black can be filled in independently from the other blanks.

In the grammar section, test-takers read 10 sentences with one blank each and choose one of the four alternatives for the blank. The following is one of the 10 questions (G2).

Test-takers read:

G2: I'll give you my spare keys in case you ( ) home before me.

A. would get B. got C. will get D. get

The correct answer:

D

Thus, the 10 items are independent from each other.

All the items were scored as either 1 (correct) or 0 (incorrect). In other words, dichotomous data were obtained. The possible total score was 30.

As for the placement decision making procedures, the cut point for the placement decision was set on the basis of the raw scores. The mean score was 19.1; the standard

deviation was 4.2. Thus, students who scored 20 or above were grouped as advanced while students who scored 19 or below were grouped as intermediate. With a few modifications due to personal scheduling problems, the students were finally classified into two levels: advanced (n = 152) and intermediate (n = 131). Then, advanced and intermediate students were randomly divided into seven classes respectively. Each class had 21 or 22 students for the advanced level and 18 or 19 students for the intermediate level. Thus, the placement decision procedures were relatively clear-cut and mechanical.

Here, it is important to note that placement decisions were made on the basis of combined scores of the listening, cloze, and grammar sections. Although the three sections may measure different aspects of L2 ability, the decision was made with the assumption that a combined score should indicate general English language ability. Thus, the placement test was considered to focus on general English language ability, not on specific skills or knowledge of English.

Analysis

As to the first research question, we examined how many items stayed within the acceptable ranges for IF (between .30 and .70) and ID (.30 or above). If the test fits the ability levels of the test-takers, it is hypothesized that we will get a larger number of items with acceptable IFs and IDs.

Regarding the second research question, we compared the original data set and the revised data set which was made by excluding the items with IFs of less than .30 or more than .70 or with IDs of less than .20. For the criterion for the revised data set, Brown (1989) used IFs between .30 and .70 and IDs of .30 or above, whereas Culligan and Gorsuch (1999) focused only on IDs (being .20 or over). First, we followed Brown's criterion. However, because, as will be shown in the results section, the number of items in the revised data set was found to be small (10 items), we took Culligan and Gorsuch's methodology into consideration. Thus, we set the ID level at .20 or over for the revised data set. If the placement decisions based on the original data set are accurate, it is hypothesized that the number of test-takers reclassified by the revised data set will be smaller.

Results

Table 1 indicates the results of the classical item analysis. In terms of IF, 13 of the 30 items (43.3%) were easier for the participants (Item facility > .70); on the other hand, 2 of the items (6.7%) were more difficult for them (Item facility < .30). The number of items with acceptable IFs between .30 and .70 was 15 (50.0% of the 30 items). In terms of ID, the number of the items with the ID value being .40 or higher was only 7 (23.3% out of the 30 items). Of the 15 items with acceptable IFs (between .30 and .70), there were 7 items with IDs of .40 or more (the very good item level); 10 items with IDs of .30 or more (the reasonably good, but possibly subject to improvement level); and 15 items with IDs of .20 or more (the marginal items, usually needing and being subject to improvement level).

Table 1

Results of Classical Item Analysis

	IF			
	IF < .30	.30 ≦ IF ≦ .70	.70 < IF	Total
ID				
ID < .20	1 (C17)	0	4 (L1/ L2/L5/L6)	5
.20 ≦ ID < .30	1 (G25)	5 (L4/L9/L10/C11/G27)	6 (L3/L7/L8/C13/C14/C15)	12
.30 ≦ ID < .40	0	3 (G21/G26/G30)	3 (C12/C20/C19)	6
.40 ≦ID	0	7 (C16/C18/G22/G23/G24 /G28/G29)	0	7
Total	2	15	13	30

The revised data set of the items with IFs between .30 and .70 and with IDs of .20 or above consisted of a total of 15 items: three listening questions (L4, L9, and L10), three cloze questions (C11, C16, and C18), and nine grammar questions (G1, G2, G3, G4, G6, G7, G8, G9, G10). Table 2 indicates the descriptive statistics of the two data sets. The paired samples t test shows a statistically significant difference between the two data sets ($t(282) = -2.74$, $p = .006$), although the two data sets show a statistically significant correlation ($r = .94$, $p = .000$). The Cronbach alpha coefficients were .71 for

the original data set and .56 for the revised data set. One possible reason for the decrease of the reliability was that the revised data set had a narrower range of item difficulty than the original data set; the revised data set excluded items that were too easy or too difficult. Even though a lower reliability coefficient was observed for the revised data set, the standard error of measurement (SEM) decreased (from 2.27 to 1.84), which in turn increases the reliability of the placement decisions, especially around the cut-score.

Table 2

Descriptive Statistics for Original Data Set and Revised Data Set (N = 283)

	Original Data Set	Revised Data Set
k	30	15
M	19.13	7.71
Max	28	14
Min	4	1
95% CI		
Lower limit	18.63	7.38
Upper limit	19.62	8.03
SD	4.21	2.78
SEM	2.27	1.84
Skewness	-0.51	0.07
SE of skewness	0.15	0.15
Z-skewness	-3.54	0.47
Kurtosis	0.10	-0.52
SE of kurtosis	0.29	0.29
Z-kurtosis	0.34	1.81
Cronbach alpha	.71	.56

As described in the participants section, the test-takers were divided into two levels of proficiency. Although the actual decision making procedures took not only the test scores into consideration but also other factors such as students' individual scheduling constraints, we used the hypothetical ideal cut-off points for the analysis of the placement decision making for this study, that is, the mean score of the test. Table 3 shows the placement results based on the two sets of data. The test-takers were assigned to the advanced or intermediate group on the basis of the mean score of each data set. Of the 283 students, 22 (7.77%) classified as advanced in the original data set

were assigned as intermediate in the revised data set; 15 students (5.30%) assigned to the intermediate group in the original data set were classified as advanced in the revised data set. Results of the McNemar test show non-significance ($p = .32$), but it is important to note that 37 out of the 283 participants (13.1%) were reassigned to different data sets.

Table 3

Discrepancies in Level Assignments Between the Original and Revised Data Sets

	Revised Data Set		
	Advanced	Intermediate	Total
Original Data Set			
Advanced	123	22	145
Intermediate	15	123	138
Total	138	145	283

Discussion

Research Question 1

The first research question asked to what extent the difficulty of the placement test fits the ability levels of the test-takers. In terms of IF, only half of the items fell within the acceptable range. Of the 30 items, 13 (43.3%) were quite easy for the test-takers; only 2 items (6.7%) were difficult. In general, this placement test can be considered to be easy for the population of this study. In terms of ID, a small number of items had good discriminatory power. The results show that the number of items with IDs of .30 or above (the reasonably good, but possible subject to improvement level) was only 13 (43.3% of the 30 items).

These results are similar to previous studies that investigated commercially produced proficiency tests. Culligan and Gorsuch (1999), for example, found that less than half of the test items they investigated had good discriminatory power. Westrick (2005) reported that only 46 out of 120 test items had acceptable IF values, and the majority of the ID values were negative. The in-house placement test examined in this study also had less than half of the items performing at acceptable levels. The implications drawn from these results are that without investigating the performance of placement test items, reliable class placement could be difficult. Furthermore,

matching the test item difficulty to the ability levels of the students is necessary in order to obtain useful information for placement purposes. The seriousness of these implications would increase for higher stakes testing situations.

Research Question 2

The second research question asked to what extent the placement decisions are accurate. The results show that, although not statistically significant, 13.1% of the test-takers were reassigned on the basis of the two sets of data. This aspect of the testing and placement procedures has serious implications concerning the validity of test score use.

As Murray (2002) pointed out, student placement into an appropriate class that matches their level is of vital importance. Misplaced students could be overwhelmed or unchallenged if placed into a course that was too difficult or too easy; therefore, placement tests must exhibit high reliability in order to accurately measure the target construct. In this regard, the reliability estimates were quite different for the two data sets, with the first data set exhibiting much higher internal consistency. However, the first data was comprised of 30 items—the revised data set contained only 15 items. Considering the number of items on both tests, a drop in reliability is to be expected. Brown (1989) had 35 items in his revised data set, but the reliability coefficient was only .63. When situated within previous studies, the reliability of the revised data set could be considered to be acceptable. Moreover, the SEM was lower for the revised data set; thus, placement decisions based on the revised data set could be more accurately made, especially around the cut-score.

This discussion suggests that classical item analysis can be utilized for immediate and long-term purposes. On the one hand, for placement decision-making on the basis of the data available at a certain time, classical item analysis may enable one to make a revised data set on which more precise decisions with a smaller SEM can be based. However, it should be kept in mind that a revised data set consists of a smaller number of items, resulting in lower reliability. The items that are not included in the revised data set should be rewritten and included on future administrations of the placement test. Thus, classical item analysis can identify those items whose functioning may not be good and provides suggestions for the revision of a placement

test. If the items performed satisfactorily, they could be included in the data set on which placement decisions would be based. Repeating these procedures would help to increase the reliability of the testing instruments and provide backing for the warrants related to test score use.

Conclusion

Classical item analysis provides information about the difficulty levels of items in relation to a sample of test-takers. Thus, the number of items with appropriate IFs indicates how well the test items fit the learners' levels of English. In addition, classical item analysis provides information about the discriminatory power of each item. IFs and IDs together suggest which items should be selected for placement decision making. Our analysis of the in-house placement test data also supported the usefulness of classical item analysis for improving placement decision making and for identifying and revising poorly functioning items.

Some limitations, however, are apparent in classical item analysis. First, as the test scores are a result of the interaction between the test and the test-takers, the results of classical item analysis are not generalizable to other testing situations or populations. This lack of generalizability makes test revision and group comparison difficult—any such comparison would lack reliability and validity. Second, placement accuracy is difficult to determine based on test scores alone. Further evidence, such as interview data, other measures of English proficiency, or test-takers' actual performance in the courses, should be examined in order to triangulate the results of the placement test. A combination of measures would produce a clearer view of placement accuracy. In practice, however, institutional and time constraints could hinder collection of other measures of proficiency (for further discussion, see Wall, Clapham, & Alderson, 1994).

Further research of placement tests should investigate the reliability and practicality of using revised data sets for placement purposes. As the number of well-performing items may be small, validating decisions based on such few items may become even more difficult. Implementing item linking based on item-response theory would be one way to overcome the problem of generalizability. Using item-response theory would also increase measurement accuracy regarding item difficulty and person

ability. Investigations of other commercially or locally produced placement tests are needed in order to shed light on tests and placement procedures that exhibit and utilize reliable measures, which in turn could provide evidence for the validity of test score use for level placement within a language program.

Notes

1. Brown (1989) examined the reliability and validity of the placement test from other perspectives than classical item analysis. For example, he administered the same test to the participants 16 weeks later and examined difference indexes to see how well the placement test fit the course content. However, in this paper, we refer to the part of Brown (1989) which relates to the question of the degree to which a particular placement test fits the ability levels of the students.
2. In addition, Brown (1989) used the value of difference index for judgment of selection.
3. In fact, like Brown (1989), Culligan and Gorsuch's (1999) study was a larger research project than described in this paper. However, we focused on the part which concerned the question of the degree to which a particular placement test fits the ability levels of the students.

References

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

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23,* 65-83.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New ed.). 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,* 7-28.

Fujita, T. (2005). *Validation of a Japanese university English language placement test.* Unpublished Doctoral Dissertation, Temple University.

Fulcher, G. (1997). An English language placement test: Issues in reliability and

validity. *Language Testing, 14*, 113-138.

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

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

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11,* 321-344.

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