Research Forum

Using Item Response Theory to Refine Placement Decisions

Greta J. Gorsuch
Texas Tech University

Brent Culligan
Aoyama Gakuin Women's Junior College

This study explores the use of Item Response Theory (IRT) or Rasch analysis in making placement decisions. The general principles underlying population-dependent classical theory standard analyses (including standard error of measure) and population-independent IRT analyses are compared and are used to point out the shortcomings of the classical analyses in making accurate placement decisions. Two sets of hypothetical cut points based on raw scores and Rasch-generated student ability estimates were applied to a set of data (N = 487) and placement decisions using the two sets of cut points were compared. Twenty discrepancies were found, meaning that five percent of the students were potentially misplaced when using their raw scores. This information may be valuable for test administrators who want to make student placements based on test results with the least amount of measurement error.

本論文では、プレースメントの決定過程での項目応答理論の利用について考察を行った。観察データを対象とする古典的理論・分析（標準誤差等）と、観察データに拘束されない項目応答理論のそれぞれの一般的原則を比較・検討し、古典的理論・分析によるプレースメント決定の問題点を指摘した。487のデータについて、素点とRaschに基づく能力推定値をもとに、それぞれのプレースメント決定を比較した結果、素点を利用した場合は5%の確率でプレースメントの判断を誤る危険性が判明した。本研究は、測定誤差を最小にしてプレースメントを実施する者に有益な情報を提供する。

In a previous study we reported on the appropriateness of the use of the SLEP test (Educational Testing Service, 1991), a commercially produced proficiency test, for placement purposes in a one-year core EFL program at a Japanese university (Culligan & Gorsuch, 1999). Using classical item analysis, we found that many test items did not discriminate well between high and low scoring students. This resulted

in a large standard error of measurement (SEM) and low test reliability (1999, p. 18). We noted that the high SEM estimate would create wide bands of score indeterminacy around program level cut points. For students with scores at or near these cut points it could be a matter of chance due to measurement error, not the students' abilities, that would put them in a higher or lower program level.

One of the positive points of classical item analysis, including item facility and index of discrimination (see Brown, 1996 for a comprehensive explanation), is that test items which discriminate effectively between high and low scoring students can be readily identified. If program administrators desire, they can score only those items, resulting in a "reduced data set" on which they could base their placement decisions. In our previous study, we demonstrated this technique with our test data and found that we could reduce the SEM and increase test reliability (Culligan & Gorsuch, 1999, p. 18). This technique will work reasonably well with programs that have administrators who are willing to use the procedure and have the equipment and trained personnel to do it.

There are two potential problems, however. First, we demonstrated that the test did not really "fit" the students who were taking it (1999, p. 17). Generally the test was too difficult, and students ended up just guessing on items. Thus many items did not offer any real information on the students' English proficiency. This "misfit" of the test to the students implies that we likely have inaccurate information about the true size of our SEM, throwing into doubt our placement decisions regardless of whether we use a full or reduced test.

Second, we pointed out previously that we live in an imperfect world. For political reasons or for reasons of timeliness and convenience, we cannot always take, or convince others to take, all the measures needed to ensure optimal student placement by scoring tests selectively. The concepts of selective test scoring and reliability, item discrimination, and SEM may be beyond the ability of concerned educators to convincingly explain to program administrators or office staff.

In this follow up study, we would like to demonstrate the use of Item Response Theory (IRT) with placement test data. We believe that an analysis offered by *Quest* 2.1, a widely available computer program in the IRT family, may give educators/administrators additional information that will enhance student placement decisions in situations where data from commercially produced proficiency tests cannot be selectively scored (Adams & Knoo, 1996).

Standard Error of Measurement Explained

The standard error of measurement (SEM) can be defined as the band of error around a test taker's score. Depending on the reliability of the test, this band of error could be several points or could be 10 or more points. If a student took the same test repeatedly, the student's scores on the tests would be normally distributed around his or her "true score." Assuming one standard deviation above or below the mean equals 34% of the distribution, the student's score would range from one SEM below the true score to one SEM above the true score about 68% of the time (34% + 34% = 68%). By extension, this means that if the student took the test 100 times, his or her score would differ from the true score by more than one SEM at least 32 times (100 - 68 = 32). On a test with poor reliability, one SEM could be 10 points. This means that a student who has a true score of 50 could go up to 60 points or down to 40 points more than 32 times out of 100 test administrations. If we look at it another way, out of 100 test takers, at least 32 students' scores are off probably by one or more SEMs. With such score variations, one can see how placement decisions based on test scores would have to take into account the SEM of the test. More importantly, by relying on a placement test with low reliability and a high SEM, we are virtually assured that some students' scores on the test will not reflect their true abilities. There is no way to determine, short of giving the test repeatedly, which students' scores are "off."

Norm Referenced SEM and What It May Not Tell You

A major problem with classical analyses of tests (of which SEM is one) is that the analyses are population-dependent. This means that test reliability, SEM, and standard deviation are a function of the number of students who took the test, as well as their scores and the distribution of their scores. In many test score distributions, the test will not be as reliable for scores that are at the middle of the distribution as for those scores at the extreme ends (high or low), "hence, the assumption of equal errors of measurement for all examinees is implausible" (Lord, cited in Hambleton, Swaminathan, & Rogers, 1991, p. 4). In other words, depending on where the students are in the score distribution of their group, they may not have the same SEM as students in other parts of the distribution. This means that wherever we create cut points for different levels in the program, we have varying levels of looseness around students' scores clustered around those cut points. Thus what SEM will *not*

tell you is the actual band of error around scores at different points in the distribution.

Item Response Theory: An Alternative to Population-Dependent Analyses

Analyses generated by Item Response Theory (IRT) have been designed to overcome the limitations imposed by population-dependent test analyses. IRT is based on the probability of a student with a given ability correctly answering a test item with a given difficulty. According to IRT, a student with high ability should have a good probability of getting an easy item correct while a student with low ability should have a poor chance of getting a difficult item correct. By feeding students' responses on all items of a test into an IRT computer program and then analyzing them along the lines of IRT, we are given estimates based on probabilities for each student's ability and each test item's difficulty. These estimates can then be applied to any student, past or future, who took or may take the test. The advantages of this will become apparent below.

In Rasch analysis, a type of IRT, indices for both the abilities of the students and the difficulties of test items are generated based on probabilities calculated by an IRT program such as *Quest* 2.1 (Adams & Knoo, 1996), which was developed for use by the Australian Council for Educational Research (available through Assessment Systems Corporation, 2233 University Avenue, Suite 200, St. Paul, MN, 55144-1629, USA). In this analysis student abilities and item difficulties are both put on the same mathematical scale, which allows student abilities and item difficulties to be directly compared. The scale typically ranges from +3 for high student abilities and difficult items to -3 for low student abilities and easy items. A student with an ability estimate of "1" will have a 50% chance of responding correctly to an item with a difficulty estimate of "1." However, a student with an ability estimate of "2" will have a 73% chance of responding correctly to an item with a difficulty estimate of "1" while a student with an ability estimate of "3" will have an 88% chance. It is the difference between ability and difficulty estimates that determines the probability of answering correctly (see McNamara, 1996, p. 166).

The hypothetical model of student abilities and item difficulties that Rasch analysis creates based on the original data is thought to hold for all students who take the test in the future. Students who subsequently take the test and are estimated by the model to have an ability level of "1" will, like the original test takers, have a 50% chance of getting items on the test with a difficulty level of "1" correct. Because the model can

be applied to subsequent test takers without regard to the number and scores of other test takers in the group, Rasch analysis is really a kind of population-independent test analysis.

Individual Measurement Error

Using a Rasch analysis of test data, we can obtain two important pieces of information that we cannot get from using classical population-dependent analysis of a test: (a) the student ability estimate and (b) the ability estimation error. The student ability estimate is created for each student by focusing on the individual student's responses on test items that tell the most about their ability. Recall that items that are too easy or too difficult for students really do not offer any information about their abilities. Students will answer easy items correctly without much thought and will usually guess at the answers to difficult items. IRT programs create a probabilistic estimate of a student's ability based on items at the point of difficulty where a particular student is not easily answering items correctly or struggling and guessing at answers. As McNamara (1996) wrote, "items have the greatest power to define the ability of the candidates in the range of ability which matches the difficulty of the item" (p. 167). The SEM, on the other hand, uses information from all students' responses to all items in the test. SEM is calculated using items that tell us very much, and very little, about students' abilities. Thus, the IRT student ability measure is a more accurate account of the true score of the student.

The ability estimation error differs from classical SEM theory in that an error estimate is created for each student ability estimate taking into account only the student's responses on the test items that are used to determine the student's ability estimate, that is, items that give us the most information about the student's ability. Both the student ability estimate and ability estimation error afforded by Rasch (IRT) analysis result in a more accurate estimate of individual students' abilities and the degree of error of this estimate. This is especially true for tests where many items are well above students' abilities and their random guesses contribute a great deal of error to the total scores.

Research Focus

In this study we are interested in whether we can refine our placement decisions by generating more information on individual students' abilities using Rasch (IRT) analysis. In particular, we want to improve placement decisions at the points where students' scores are clustered

around hypothetical program-level cut points. We want to know if individual students' ability scores as provided by Rasch analysis indicate that students clustered around hypothetical program cut points have been placed into the wrong program levels.

Method

Participants

Only a brief description of the participants will be given here. For a full description, see Culligan and Gorsuch (1999). The participants in this study were 487 first year students at a private Japanese university near Tokyo. This number is well above the minimum of 100 students initially needed to complete Rasch analysis. They were predominantly Japanese, were eighteen years of age, and were liberal arts majors. Around 80% of the subjects were male.

Materials

A full description of the SLEP test form (Educational Testing Service, 1991) used in the study appears in Culligan and Gorsuch (1999). Briefly, the SLEP test is a 150-item measure of English proficiency normed on non-native English-speaking secondary school students in the U.S. It includes listening and reading subsections.

The computer program used in this study is *Quest* 2.1 for Macintosh computers (Adams & Knoo, 1996). It uses a single parameter Rasch measurement model and can provide analyses on both test items (items) and test takers (cases). Because *Quest* 2.1 is actually a FORTRAN program adapted for use with a Macintosh, it does not make use of the dialog boxes Macintosh and Windows users are familiar with. Instead, highly defined, non-intuitive commands must be typed in to create the analyses desired. In this study, we have given the precise commands we used to conduct our analysis. We hope this will help readers conduct their own Rasch analyses.

Procedure

In April 1996, SLEP test data for 487 students was read by an optical scanner and entered into a spreadsheet program. To prepare the data for analysis using *Quest* 2.1, the data was converted into tab-delimited text and pasted into a word processing program document. In order for the program to accurately "read" the data, the spaces created by the tabs were then eliminated using a search/replace function in the word processing program. This created a data set that looked like the data in Table 1.

Table 1: Sample Data Set

96122011100010101000 . . .
96130100011101010100 . . .

In the full data set, the 1's and 0's go off to the right and each line of data goes down to line 487. Note that there are no spaces between the characters. The first six numbers were student ID numbers (the numbers used here have been fabricated) and the following 150 "1" and "0" characters on each line indicates the students' correct and incorrect responses to each item. The data set was given the name slep.dat and was placed directly into the *Quest* 2.1 folder in the computer. The following batch commands were typed into a word processing program and the program was saved as slep.ctl and placed in the *Quest* 2.1 folder. We have given the purpose of each command in italics (see Table 2).

Table 2: Batch Commands in slep.ctl

Command	Purpose of Command
Title SLEP Pretest	*Gives a running header for the program output.*
data_file slep.dat	*Tells the program which Data Set to use.*
format name 1-6 items 7-157	*Tells the program which characters in the Data Set should be analyzed.*
estimate	*Tells the program to analyze the data.*
show>>out1.txt	*Gives test reliability, summary of fit indices, and an item/case map.*
show cases ! order=estimate >> out3.txt	*Requests the program to show the student ability and student ability error estimates for all cases (students), rank student ability estimates in descending order, and to put the information into a document called out3.txt, which you can open after quitting Quest 2.1.*
quit	*Instructs the program to quit.*

Note. The commands on the left would ordinarily appear single-spaced. Blank lines have been added in this table to correspond to the descriptions of the purposes of the commands.

To run the analysis, we launched the *Quest* 2.1 program and typed in: submit slep.ctl. The program completed the analysis and put the results into the out3.txt document we specified.

Data Analyses

In order to generate hypothetical student placement cut points, descriptive statistics for students' raw scores were calculated using *Microsoft Excel* 5.0 (1985-1993). The raw scores are what most program administrators would use to calculate the mean, standard deviation, and SEM of the data in non-IRT analyses. The raw scores were rounded to the nearest whole number. We used a raw score of 70 as the mean, 82 as the upper cut point, and 57 as the lower cut point. Assuming we wanted to place students into three groups (advanced, intermediate, beginner), students with a raw score of 82 or above would be placed in the advanced group, students with raw scores ranging from 58 to 81 would be placed in the intermediate group, and students with raw scores of 57 or lower would be placed in the beginners group.

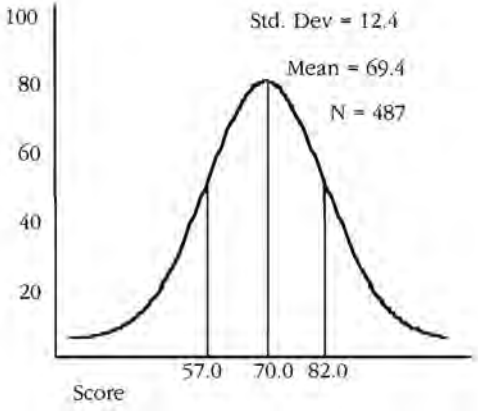
In order to match the raw scores to the equivalent Rasch-generated student ability estimates, we looked at the *Quest* 2.1 output in the "score" column for all students who scored at the lower cut point and the equivalent student ability estimates. We identified a common ability estimate equivalent to the upper and lower cut point scores on the SLEP test (see Table 3). For a visual representation of the hypothetical cut points plotted on the score distribution, see Figure 1.

Table 3: Descriptive Statistics of SLEP Test Data and Equivalent Student Ability Estimates (N = 487, 150 items)

Statistic	Raw Score	Rounded	Ability Estimate (Rasch)
M	69.36	70.00	-.13
SD	12.38		
Upper cut point	81.74	82.00	.26
Lower cut point	56.98	57.00	-.57

We then looked at the data to identify those students with discrepancies, where their raw score suggested they should be in one level (advanced, intermediate, beginner) but their Rasch-generated student ability estimate placed them in another. Recall that the data was sorted by *Quest* 2.1 according to student ability estimates in descending order

Figure 1: Hypothetical Cut Points on Test Score Distribution



(Table 2). We identified students whose raw scores were below the higher raw score cut point but whose ability estimates were above the student ability estimate cut point. For example, a student, such as case 1069 (see Table 4 for sample *Quest* 2.1 output), with a raw score of 75 would be placed in the intermediate level, but based on his or her student ability estimate of .33, would be placed in the advanced group. We repeated the procedure for the lower cut point.

Table 4: Sample *Quest* 2.1 Output

Case Estimates In Estimate Order All on All (n = 487, L = 150, Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
43 1043	95	129	1.13	.22	.94	.85	-.57	-.66
225 1225	107	150	1.13	.19	.85	.70	-1.83	-1.77
69 1069	75	127	.33	.20	.81	.73	-2.62	-1.97

As shown in Table 5, there were a number of discrepancies between placement decisions using raw scores and decisions using Rasch-generated student ability estimates. Using the raw score cut points described in Table 3, 74 students were placed in the advanced level, 333 in the intermediate level, and 80 in the beginner level. However, using the Rasch-gener-

ated student ability cut points, 82 students were placed in the advanced level, 337 in the intermediate level, and 68 in the beginner level.

We identified 20 students whose raw scores would place them in one level but whose student ability estimates would place them in another. Eight students were found whose ability estimates placed them in the advanced group, while their raw scores placed them in the middle group. Twelve students who would have been placed in the beginner group based on raw scores were placed in the intermediate group based on student ability estimates.

Table 5: Discrepancies in Student Placement

Students Placed Using Raw Scores			
Level:	Advanced	Intermediate	Beginner
Score:	82 and above	81 to 58	57 and below
Number of Students:	74	333	80
Students Placed Using Rasch Ability Estimates			
Level:	Advanced	Intermediate	Beginner
Score:	.26 and above	.25 to -.56	-.57 and below
Number of Students:	82	337	68

Discussion

In this study, we attempted to refine our placement decisions by obtaining more information using Rasch (IRT) analysis. We found there were 20 discrepancies between student placement using their raw scores and their ability estimates generated by IRT analysis, meaning that 20 students in this hypothetical situation were potentially misplaced (5% of all test takers in the group). We therefore suggest that test administrators could use this procedure to identify such students. We also suggest that test administrators should investigate which scoring method, raw scores or Rasch student abilities, is the best predictor of group membership for their situation. Such an investigation would involve collecting longitudinal data on students' progress and ultimate achievement in their classes, as well as administrative procedures to identify misplaced students and reassign them once the program has started. While an IRT analysis is not a substitute for an in-depth analysis and development of placement tests and placement procedures, IRT can be used by program administrators

to make the best out of a less than ideal situation.

While the results of this study cannot be generalized to other schools that use the SLEP test, the tools outlined in this paper can be applied to all situations involving tests where there are 100 or more test takers. We urge educators to use IRT in making placement decisions, and then to report the successes and challenges of doing so in real life programs. Of particular interest would be reports on the use of IRT in conjunction with longitudinal data to investigate whether the Rasch model of student ability and item difficulty estimates based on an initial group of test takers held for subsequent test takers with much higher or lower levels of ability.

Acknowledgments

We would like to thank Steve Ross for his assistance on Item Response Theory and the finer points of using Quest 2.1. *We would also like to thank Dale T. Griffee and two anonymous* JALT Journal *reviewers for their helpful comments on an earlier draft of this article.*

Greta Gorsuch is an Assistant Professor in the Department of Classical and Modern Languages at Texas Tech University. She is interested in testing, performance assessment, and research methodology.

Brent Culligan is a full time instructor at Aoyama Gakuin Women's Junior College. A doctoral candidate at Temple University Japan, he is interested in second language vocabulary acquisition and testing.

References

Adams, R.J., & Knoo, S.T. (1996). *Quest* (Version 2.1) [Computer software]. Victoria, Australia: Australian Council for Educational Research.

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

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(1), 7-28.

Educational Testing Service. (1991). *SLEP test manual*. Princeton, NJ: ETS.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

McNamara, T. (1996). *Measuring second language performance*. Harlow, UK: Addison Wesley Longman.

Microsoft Excel 5.0 (1985-1993). Seattle, WA: Microsoft Corporation.

(Received September 16, 1999; revised February 26, 2000)