SHIKEN

Volume 19 • Number 1 • April 2015

Contents

JALT

Testing and Evaluation SIG

Foreword

Trevor A. Holster
trevholster@gmail.com
TEVAL SIG Publication Chair

Welcome to the first issue of *Shiken* for 2015. In this issue, Koizumi et al. investigated regression to the mean (RTM) and the standard error of difference (SED), issues which are surprisingly underrepresented in TESOL literature. Although Koizumi et al. and Swinton (1983) demonstrated RTM using TOEFL scores, this problem will affect any pretest-posttest comparisons, whether for research studies or for monitoring student learning. As both Swinton and Koizumi et al. made clear, RTM can lead to misinterpretation of test results, invalidating research findings and compromising instructional decisions, so the lack of awareness of the problem is worrisome. Swinton's study, published as an official ETS Research Report, recommended using two pretests to estimate and correct for RTM, so Koizumi et al.'s guidance on how to estimate RTM when only a single pretest is available provides a much more practical approach for the situation that most teachers and program administrators face.

Smiley, who will be familiar to many TEVAL members as the coordinator of the JALT Materials Writers SIG, provided an account of the difficulties that teachers face in developing skills in test analysis. Smiley's use of Microsoft Excel to conduct classical test theory (CTT) analysis was highly commendable and the guidance provided by books such as Brown's (2005) *Testing in Language Programs* is more than sufficient for the needs of teachers who need to analyze classroom tests. However, Smiley needed to criterion reference test questions against curriculum objectives and textbook content, a purpose for which Rasch analysis is ideally suited (Brown & Hudson, 2002). Linacre's (2014) Winsteps software package provides for quick and detailed Rasch analysis of overall test performance and individual items, but Smiley also cautioned that novices may be discouraged by the steepness of the learning curve involved in learning Rasch analysis.

J.W. Lake and I responded to Smiley's article by highlighting two key features of the Rasch model: the Wright map and the assumption that all items discriminate equally. Our aim was not to provide groundbreaking new insights, but rather to demonstrate that Rasch analysis can provide information to guide instructional decisions that is not easily available from CTT analysis and that Rasch results can be presented in graphical forms that are conceptually simple enough that novices can interpret them without requiring extensive technical training.

Finally, J.D. Brown's regular *Statistics Corner* column reviewed the range of techniques and analyses that have been used in the testing of intercultural pragmatics ability. Pragmatics, which deals with the relationship between context and meaning, is crucial to language proficiency, and thus to assessment, evidenced by the growing body of research on its assessment documented by Brown. Hopefully the inclusion of pragmatic features in assessment will result in positive washback, where textbooks and classroom instruction reflect the testing of intercultural pragmatics. One point that stands out about Brown's review is the increasing sophistication of the analyses used in testing intercultural pragmatics ability, which is evidence of the complex nature of the interaction between language and context. In particular, the increasing use of Facets analysis to account for rater effects (see McNamara, 1996, for an accessible introduction) raises questions about how to incorporate pragmatics into classroom assessments because teachers frequently act as interlocutors and/or raters. Given that pragmatics is concerned with what is appropriate in different contexts and when faced with different interlocutors, the elicitation of pragmatics performances in a classroom by a teacher raises questions of how to interpret the results, i.e. the construct validity of the assessment. The problematic nature of authenticity in classroom contexts is well recognized (see van Lier, 1996, for example). Facets analysis, which isolates contextual variables as *facets* of a performance, can address some of these concerns, but the complexity of the analysis often makes the findings incomprehensible to non-specialists, as Smiley's article in this issue reported. This doesn't preclude positive washback from tests of intercultural pragmatics, but it does raise questions about what degree of assessment literacy teachers need for positive washback to occur.

The TEVAL SIG has been working for many years to make technical issues more accessible to classroom teachers through J.D. Brown's *Statistics Corner,* Jim Sick's series of articles on Rasch analysis, and Tim Newfield's articles on assessment literacy, but the articles by Smiley and Brown are important reminders of the need for workshops and introductory articles aimed at novice language testers. The JALT Pan-SIG2015 conference will be held in Kobe on the weekend of 16-17 May, 2015. Many of our officers and members will be attending, so this is an excellent opportunity to see the work of TEVAL SIG members and the members of other JALT SIGs, and to raise any questions or concerns about testing and assessment. We look forward to seeing you at Pan-SIG2015.

References

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New Ed.).* New York: McGraw-Hill.

Brown, J. D., & Hudson, T. D. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Linacre, J. M. (2014). Winsteps (Version 3.81.0). Retrieved from http://www.winsteps.com

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Pearson Education.

Swinton, S. S. (1983). A manual for assessing language growth in instructional settings. Princeton: Educational Testing Service.

van Lier, L. (1996). *Interaction in the language curriculum: Awareness, autonomy and authenticity*. Harlow: Pearson.

Call for Papers

Shiken is seeking submissions for publication in the November 2015 issue. Submissions received by 1 September, 2015 will be considered, although earlier submission is strongly encouraged to allow time for review and revision. *Shiken* aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference

Rie Koizumi[1], Yo In'nami[2], Junichi Azuma[3], Keiko Asano[1], Toshie Agawa[1], and Derek Eberl[1]
rkoizumi@juntendo.ac.jp
1. Juntendo University
2. Chuo University
3. Kobe Gakuin University

Abstract

Regression to the mean (RTM) and the standard error of difference (SED) are two artifacts commonly observed in pretest–posttest designs, but they are rarely addressed in practice. We examined whether second language (L2) learners' change in scores reflected change in their L2 proficiency, by investigating whether their actual scores exceeded those that considered RTM and SED; we did so by using pretest–posttest data of the Test of English as a Foreign Language Institutional Testing Program (TOEFL ITP) at a Japanese university across three years. We found moderate degrees of RTM, but also found that more than one-third (33.33–46.03%) of students increased their scores beyond RTM and the SED. We discuss the importance of considering RTM and SED in analyzing pretest–posttest data.

Keywords: considering errors in practice, pretest–posttest data, TOEFL ITP, Japanese university students

Change in pretest and posttest scores is often investigated using descriptive statistics such as means, or statistical significance tests such as paired *t*-tests. However, such change is subject to many factors other than change in true ability. Examples of other factors are maturation, history, and practice effects (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). Thus, analysis in change requires careful examination beyond descriptive statistics and/or statistical significance tests, before we can confidently conclude that a change in ability was indeed observed.

Regression to the mean (RTM) and the standard error of measurement (SEM) are two factors that affect pretest–posttest score changes and are commonly observed in language testing. RTM refers to a situation where pretest scores farther from the mean are probabilistically likely to cluster around the posttest mean. Thus, students who scored much lower than the pretest mean tend to increase their scores in the posttest more than those who scored a little lower than the pretest mean. Alternatively, students who scored much higher than the pretest mean tend to lower their scores more than those who scored a little higher than the mean.

Another consideration in pretest–posttest designs is that every test score includes measurement error, often known in practice in the form of the standard error of measurement (SEM). SEM refers to "the standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions"; it is "usually estimated from group data" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, pp. 223–224). For example, when a test taker obtains a score of 500, due to the SEM, his or her true score could be 495 or 510. This suggests that these three scores are all within the margin of error and that seemingly different scores are mere artifacts owing to measurement error. SEM is often used to interpret a single score or to compare two scores from the same test. When we compare two scores from the same test or parallel tests administered in different occasions to the same person, we can instead use a special version of the SEM that is tailored for repeated measurements. This is called the standard error of difference, and it refers to the measurement of error for scores obtained from two test administrations (Harvill, 1991).[1] The standard error of difference is typically abbreviated as SE_{diff} or SED—the latter of which we use in this paper.

Although RTM and SEM (and SED) are frequently observed in language testing and widely known in the measurement literature, they seem to have been rarely considered in practice when interpreting score change. We will report on how to consider RTM and SED; we then illustrate this, using second language (L2) pretest–posttest data from the Test of English as a Foreign Language Institutional Testing Program (TOEFL ITP). For this purpose, we will first explain RTM and SED.

Regression to the Mean (RTM)

According to Campbell and Kenny (1999), many phenomena unrelated to language tests can also be explained by RTM. For example, tall parents tend to have children who are not as tall as their parents (i.e., children's height moves nearer to the children's mean); sports rookies who fared exceptionally well in their first year cannot be as successful in the following year (i.e., their performances regressed to those of the average players); and a sequel to a hit movie is less popular than the earlier one (i.e., the sequel performances became nearer to the mean of movies).

RTM is prevalent in language test data involving pretest–posttest or multi-time-point scores. Thus, an increase in pretest–posttest scores cannot be automatically attributed to the beneficial effect of treatment or to the actual growth of ability. Given that gain-score studies are widely used and yet their value could be undermined by RTM, many studies have proposed how to address RTM in psychology, education, medicine, and other fields (Campbell & Kenny, 1999).

There are several ways to deal with RTM, and they are categorized as those applied before data are collected and those after data are collected (Campbell & Kenny, 1999). The former method is recommended, as it allows one to design a study with RTM in mind.

There are three pre-data-collection methods. First, researchers should examine research designs carefully. For example, this includes (a) assigning the same number of participants to each group randomly in experimental/control groups designs, (b) avoiding pretest–posttest designs that have only experimental groups and no control groups, and (c) always conducting one or more pretests. A second approach is to consider the plausible effects of a third variable that affects the study findings and to model them (for example, as a covariate) when designing a study. For example, if experimental and control groups widely differ in terms of English-learning motivation, the pretest–posttest difference could be due to the treatment, the difference in their motivation, or both. Modeling a third factor—motivation, in this case—could reduce both the SEM and RTM. A third way is to use measures and tests that feature high reliability. The higher the reliability of a test is, the more consistently it measures knowledge and ability. This leads to a decrease in SEMs (e.g., Bonate, 2000). Concurrently, researchers should use tests that are shown to be parallel in content and difficulty and highly correlated when they are administered at the same time or on close occasions. This is because RTM occurs when two tests do not correlate perfectly. The more highly two tests correlate, the less RTM is observed; in contrast, a weaker intercorrelation leads to greater RTM.

In addition to these pre-data-collection methods for addressing RTM, there are several ways of dealing with this issue, even after data are collected—at the data analysis stage. We will present them in the Method section, below. All these pre- and post-data-collection methods are extensively discussed in Campbell and Kenny (1999, Chapter 10). While Marsh and Hau (2002) report that RTM cannot be completely controlled for, even with the latest analytical method of multilevel analysis, the pre- and post-data-collection methods we describe in this paper are still useful in even partially addressing the regression effect.

Despite the prevalent possibility of RTM effects in pretest–posttest designs, they have not been adequately addressed in L2 growth-assessment studies. One exception is Swinton (1983), who aimed to offer guidelines for establishing a language-growth benchmark at local institutions. His method was based on the recommendations of Cronbach and Furby (1970), and is as follows: administer three tests on the same examinees—namely, Pretest A at the beginning of a semester, Reliability Test B one week after Pretest A, and Posttest C at the end of the semester; develop a regression equation relating Reliability Test B to Pretest A, in order to estimate score change due to nonintervention (e.g., measurement error and practice effects); and examine whether a Posttest C score is higher than that expected from the regression equation. Suppose our regression equation is: Test B = 200 + .7*(Pretest A), and a student scores 100 on Pretest A and 150 on Test B. Using the regression, the expected score on the Posttest C (when RTM occurs and there is no actual gain) is 270 (= 200 + .7*[100]). If the actual score on the Posttest C is 280—that is to say, greater than 270—it indicates growth in ability. The gain is 10 (= 280 – 270), not 180 (= 280 – 100).

Perhaps the most comprehensive recent study on RTM is that of Marsden and Torgerson (2012). Using the database *Educational Resources Information Centre*, they conducted a methodological review of single-group, pretest–posttest empirical studies published in 13 educational research journals in 2009. Of 490 studies published in that year, 64 (13%) were found to have evaluated innovative interventions and used experimental, quasi-experimental, or pre-experimental designs. After excluding 48 studies that did not meet their inclusion criteria (e.g., studies that did not "have at least one quantified measure," p. 588), Marsden and Torgerson had 16 studies left; they report that none of the study authors described the potential effect of RTM, although other potential factors (e.g., maturation and time) were mentioned. This indicates that RTM is not widely recognized among scholars, and that a better understanding of RTM could lead to improvements in the interpretation of study findings.

Standard Error of Difference (SED)

Test scores include errors caused by variations in the content and format of test items and whole tests, and by inconsistencies in test administration and scoring (see, for example, Fulcher, 2010; Hughes, 2003). These causes could lead to larger errors and lower test reliability, whereas tests that have high reliability have smaller test-score errors.

Equation1 below (Harvill, 1991, p. 186, Formula 10) is used to calculate SEDs—that is, the measurement of errors for scores obtained from two administrations (Harvill, 1991). The value shows the degree to which a test score changes at 68% probability because of errors under two test administrations. In other words, it shows the 68% probability of score variation. Equation 1 indicates that tests with larger standard deviations (*SD*s) have a larger SED. Further, larger SED values indicate larger error, resulting in test scores featuring greater fluctuation (see, for example, Carr, 2011, on how to calculate *SD* and reliability). Greater confidence in score variation can be obtained by calculating the 95% probability of score variations, using Equation 2 (Harvill, 1991, p. 186); however, we use only Equation 1 because "score bands which are 68 percent confidence intervals … are most commonly used in practice" (p. 184).

SED (for 68% probability in comparing two scores from the *same* test taker[2])
= (*SD* of the pretest) * (√[2 – (Reliability of the pretest) – (Reliability of the posttest)]) (1)

SED (for 95% probability) = 1.96 * SED (for 68% probability) (2)

Suppose that the SED (for 68% probability) of the TOEFL ITP is 15. This would mean that under two test administrations, this test score can fluctuate by 15 at the probability of 68%. When a pretest score is 480 and a posttest score is 520, the error range of the pretest score is between 465 and 495; 520 is not included in this range. Thus, the pretest and posttest scores are highly likely to differ, and we can assert this with confidence. However, if the posttest score is 485, it would be difficult to argue that the two test scores differ.

As with RTM, SEDs are not always reported or considered during score interpretation. Good practices are seen in the *TOEIC User Guide* (Educational Testing Service, 2007) and the *TOEIC Examinee Handbook* (Educational Testing Service, 2008), both of which explain that the 68% probability SED for each of the TOEIC listening and reading sections is 35. If a student's listening score improves from 300 to 340, this indicates real growth in that student's listening proficiency, as the score of 340 lies outside the 265–335 SED range.

The Current Study

We address three research questions (RQs) to examine whether RTM is observed (RQ1), and to what degree RTM and the SED affect findings (RQ2 and RQ3) in using the TOEFL ITP. The three RQs are as follows.

1. Is there any evidence of the regression to the mean (RTM) in pretest–posttest data?
2. What percentage of students increased or decreased their scores beyond RTM?
3. What percentage of students increased or decreased their scores beyond the SED?

The results could contribute to our understanding of how to separate students' real growth in proficiency from RTM and SED. This would, in turn, strengthen arguments regarding whether or not students had increased in ability.

Method

Participants and Instrument

We used data from first-year students at the tertiary level who took the TOEFL ITP (Level 1) twice—in April and December—at a private university in Chiba, across three years (n = 120 in 2012; n = 125 in 2013; n = 126 in 2014). The TOEFL ITP was conducted to assess growth in students' L2 English proficiency, and to place students into five English classes in the subsequent year. Additionally, each student needed to obtain a TOEFL ITP of 475 or higher, or a TOEFL Internet-based test (iBT) of 53 or higher, to advance to the second year. Thus, students were generally motivated to study hard to meet the requirement. The test was also conducted to evaluate the effectiveness of the English program.

The TOEFL ITP is designed to assess the English proficiency of nonnative speakers. It has three sections, all in paper-and-pencil multiple-choice formats; it consists (in order of appearance) of a listening section (50 items, 35 minutes), grammar section (40 items, 25 minutes), and reading section (50 items, 55 minutes). A total score ranges from 310 to 677, with an SEM of 13 (Educational Testing Service, n.d.). Each student receives score reports that show each of the three section scores and the total score.

Analyses

The TOEFL ITP total scores from the April and December administrations over the three-year period were used. Of the various methods available for analyzing the degree of RTM, we utilized two methods that we consider the most accessible. First, to address RQ1, we correlate the change scores (posttest minus pretest) and the pretest scores. If there is an RTM effect, we see a substantial and negative correlation. The higher a negative correlation is, the higher the degree of RTM will be (e.g., Marsden & Torgerson, 2012; Roberts, 1980; Rocconi & Ethington, 2009; Rogosa, 1995). The rationale here is that negative correlations are derived from lower pretest scorers who are likely gaining a higher posttest score and from higher pretest scorers who are likely gaining a lower posttest score.

While the first method is group-based and produces a single value that shows the overall extent of RTM, the second method is individual-based: It calculates expected individual posttest scores while assuming RTM, and compares them to the actual scores. Using Equation 3 below (Campbell & Kenny, 1999, p. 26), we calculated an expected posttest score per person and compared it to his or her actual posttest score, while bearing in mind RQ2. We calculated the percentage of students with actual posttest scores that were higher than their expected posttest scores—that is, the percentage of those whose growth exceeded RTM. Table 1 shows how our three-year data were applied to the Equation.

$$\text{Expected posttest score} = M_{\mathrm{y}} + r_{\mathrm{xy}}(SD_{\mathrm{y}}/SD_{\mathrm{x}})(\mathrm{X} - M_{\mathrm{x}}) \tag{3}$$

Table 1

Expected Posttest Scores and Actual Posttest Scores

x	2012	2013	2014
M_y = posttest score mean	508.87	522.82	539.76
r_{xy} = correlation between pretest and posttest scores	.79	.83	.80
SD_y = standard deviation of posttest scores	34.17	42.58	36.16
SD_x = standard deviation of pretest scores	40.50	47.34	44.97
M_x = pretest score mean	507.13	508.01	510.51
X = Actual pretest score of a student (example)	500	517	557
Expected posttest score (example)	504	530	570
Actual posttest score (example)	510	547	560
	(gain)	(gain)	(no gain)

For example, for a student in 2013 with a pretest score of 517, his expected posttest score was 530 (522.82 + .83*[42.58/47.34]*[517 – 508.01]); the actual posttest score was 547. This student's actual posttest score was larger than the posttest score forecast by RTM; this suggests that the score gain reflects improvement in his ability, rather than RTM. It should be noted that the standard deviations of the posttest scores were all smaller than those of the pretest scores; this indicates a narrower distribution of posttest scores and may serve as one piece of evidence of RTM.

To examine RQ3, Equation 1 was used; SEDs for 68% probability were calculated as follows.

SED (for 68% probability in comparing two scores from the same test taker)
= (Standard deviation of the pretest) * (√[2 – (Reliability of the pretest) –
(Reliability of the posttest)]) (1)

SED in 2012 = (40.50)*(√[2 – (.96) – (.96)]) = (40.50)*(0.28) = 11.46
SED in 2013 = (47.34)*(√[2 – (.96) – (.96)]) = (47.34)*(0.28) = 13.39
SED in 2014 = (44.97)*(√[2 – (.96) – (.96)]) = (44.97)*(0.28) = 12.72

We used the reliability of .96 for the TOEFL ITP, as reported by Educational Testing Service (n.d.), as the reliability index for both the pretest and posttest.[3] We calculated the percentage of students who had actual posttest scores higher than the SED—that is, the percentage of those whose growth exceeded the SED. All analyses were conducted using the *Comparing paired samples* (which includes paired *t*-tests) and *Correlation* pages in the langtest.jp Web App (Mizumoto, n.d.). The app runs on several well-known R packages and produces various useful figures based on data pasted into its website's designated space.

In summary, we will (a) examine correlations between change scores and pretest scores (for RQ1), (b) calculate the predicted posttest scores and the percentages of students whose posttest scores exceeded the predicted ones (for RQ2), (c) consider the SED and calculate the percentages of students whose posttest scores exceeded the SED (for RQ3), and (d) synthesize the findings from (b) and (c).

Results and Discussion

Data distribution

Means, standard deviations, and correlations between the pretest and posttest scores in the 2012–2014 data are presented in Table 1 above. Boxplots of pretest–posttest scores and changes in individual scores in 2012, meanwhile, are presented in the left-hand panel of Figure 1, while the right-hand panel shows clearly that pretest scores—especially extreme ones—tend to converge to the posttest mean. This is consistent with the smaller standard deviation for the posttest (34.17, compared to 40.50 for the pretest [see Table 1]). These results suggest a certain degree of RTM in the 2012 data.

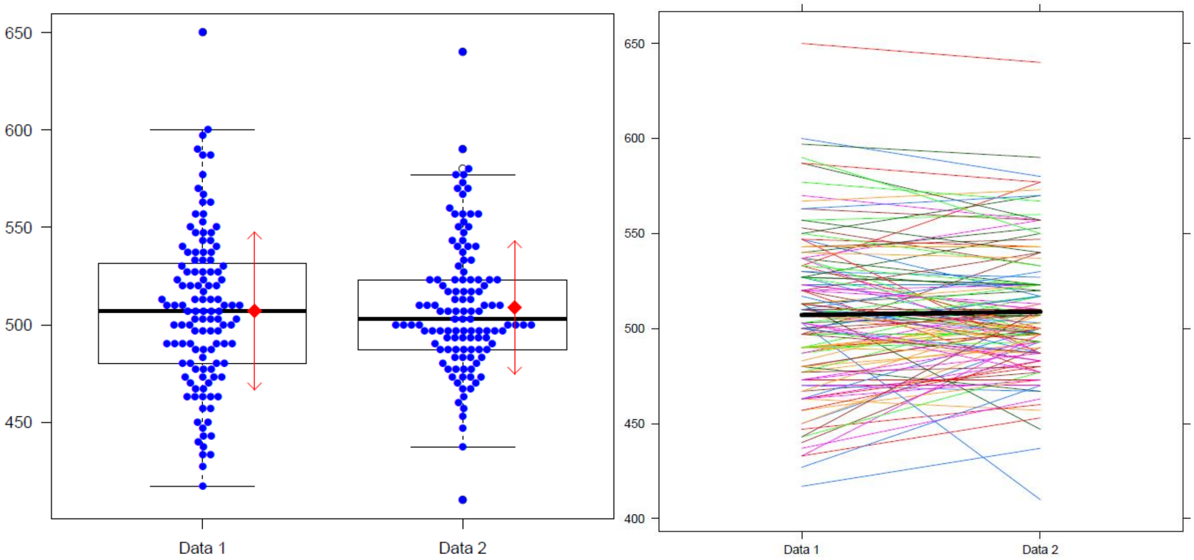

Figure 1. Left: Boxplots of pretest–posttest data in 2012. Data 1 = Pretest scores; Data 2 = Posttest scores. ±1 standard deviations are represented by arrows. See Field (2009) for interpretation of the boxplots. This note also applies to the left-hand panel of Figures 2 and 3. Right: Changes in individual scores between pretest and posttest data in 2012. Thick line indicates the mean difference. This note also applies to the right-hand panel of Figures 2 and 3.

The left-hand panel of Figure 2 shows pretest–posttest scores from the 2013 data, and the right-hand panel shows overall that extreme pretest scores were likely to converge toward the posttest mean. The posttest standard deviation was smaller than the pretest standard deviation (42.58 and 47.34, respectively). These results, again, provide some evidence of RTM in the 2013 data.

Figure 3 shows pretest–posttest scores from the 2014 data, with the right-hand panel showing overall that extreme pretest scores were likely to converge toward the posttest mean. The posttest standard deviation was smaller than the pretest standard deviation (36.16 and 44.97, respectively). Again, these results provide some evidence of RTM in the 2014 data.

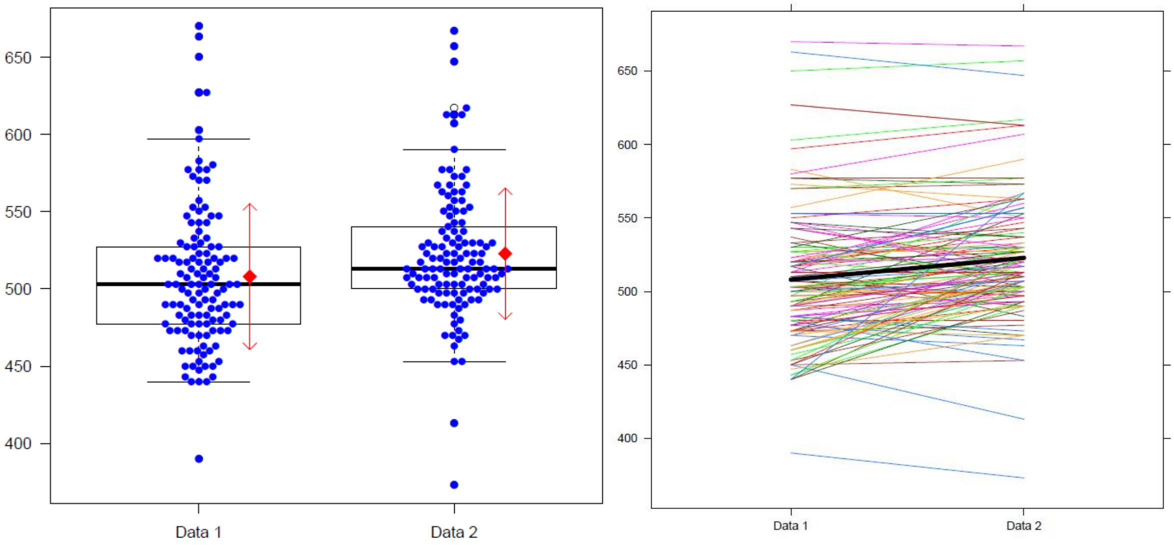

Figure 2. Left: Boxplots of pretest–posttest data in 2013. Right: Changes in individual scores between pretest and posttest data in 2013.

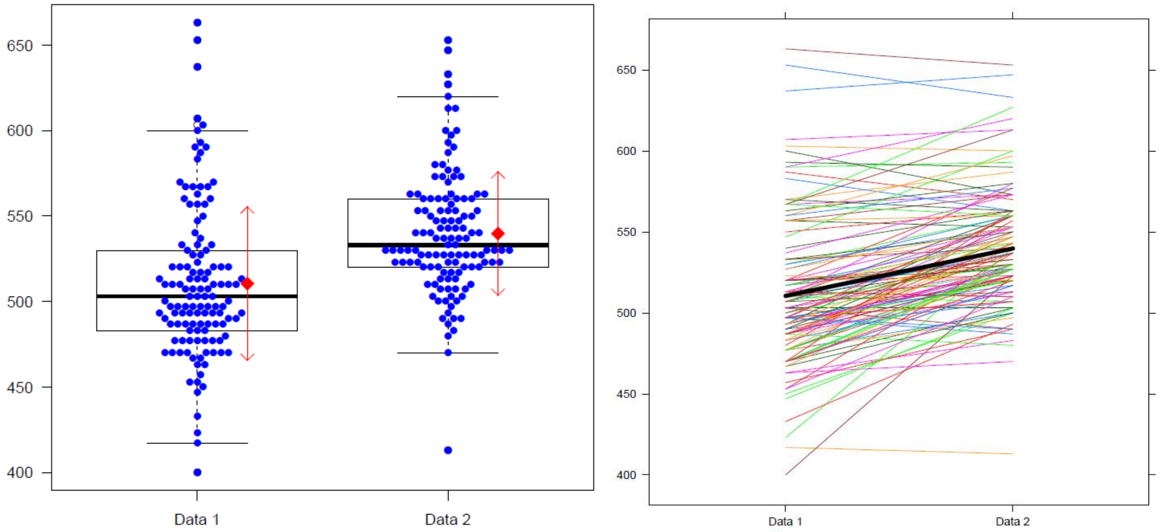

Figure 3. Left: Boxplots of pretest–posttest data in 2014. Right: Changes in individual scores between pretest and posttest data in 2014.

For reference purposes, we used paired *t*-tests to compare the pretest–posttest scores. Table 2 shows that in 2012, there was no significant difference between the two scores, with a negligibly small effect size according to Plonsky and Oswald's (2014) guideline for interpreting effect sizes (i.e., within-group contrast d = 0.6 for small, 1.0 for medium, and 1.4 for large). In 2013, there was a significant difference with a negligibly small effect size, whereas in 2014, there was a significant difference between the two scores with a small effect size.

Table 2
Results of Paired T-tests and Effect Sizes

	Paired *t*-test	Effect size
2012	*t* = -0.76, *df* = 119, *p* = .45	*d* [95% CI] = 0.05 [-0.07, 0.16], *g* = 0.04 [-0.07, 0.16], δ = 0.04
2013	*t* = -6.24, *df* = 124, *p* < .01	*d* [95% CI] = 0.32 [0.22, 0.43], *g* = 0.32 [0.22, 0.43], δ = 0.31
2014	*t* = -12.27, *df* = 125, *p* < .01	*d* [95% CI] = 0.69 [0.56, 0.81], *g* = 0.68 [0.56, 0.80], δ = 0.65

Note. *d* = Cohen's *d*; *g* = Hedges' *g*; *δ* = Glass's delta.

RQ1: Is there any evidence of the regression to the mean (RTM) in pretest–posttest data?

Figures 4–6 show scatterplots and correlations between change scores and pretest scores; the relationships therein are consistently negative and moderate (*r* = –.54, –.45, and –.59, respectively). Negative correlations indicate RTM, and the higher they are, the greater the degree of RTM (e.g., Roberts, 1980; Rocconi & Ethington, 2009; Rogosa, 1995). This is because negative correlations suggest a greater change in scores when pretest scores are lower, as well as less-positive-change scores or more-negative-change scores when pretest scores are higher. Negative and moderate correlations suggest moderate degrees of RTM. As evidenced by negative and larger correlations, the 2014 data show more serious RTM than the 2012 and 2013 data, and the 2012 data show more serious RTM than the 2013 data.

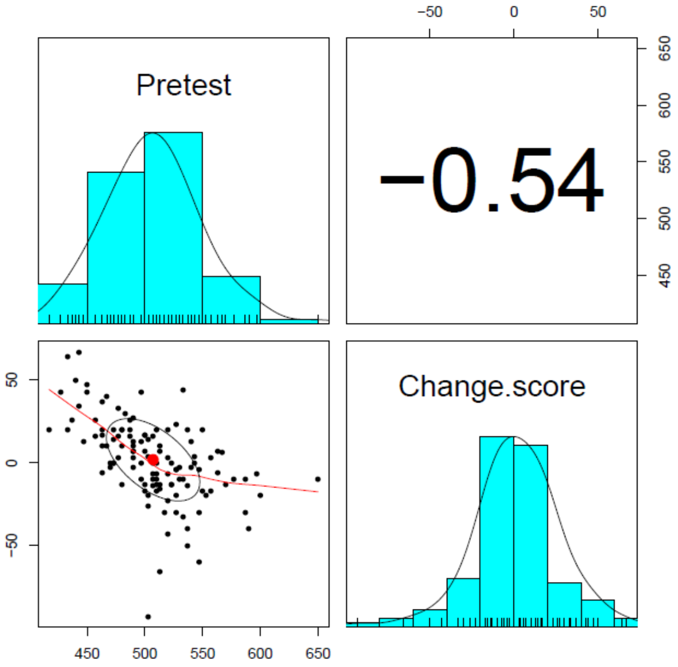

Figure 4. Scatterplot and correlations in 2012. The left bottom scatterplot has pretest scores on the X axis and pre–post change scores on the Y axis. The red curve shows loess smooths, the large black circle shows correlation ellipses, and the large red dot indicates the means of the X and Y axes (see Ogasawara, 2014; Revelle, 2014).

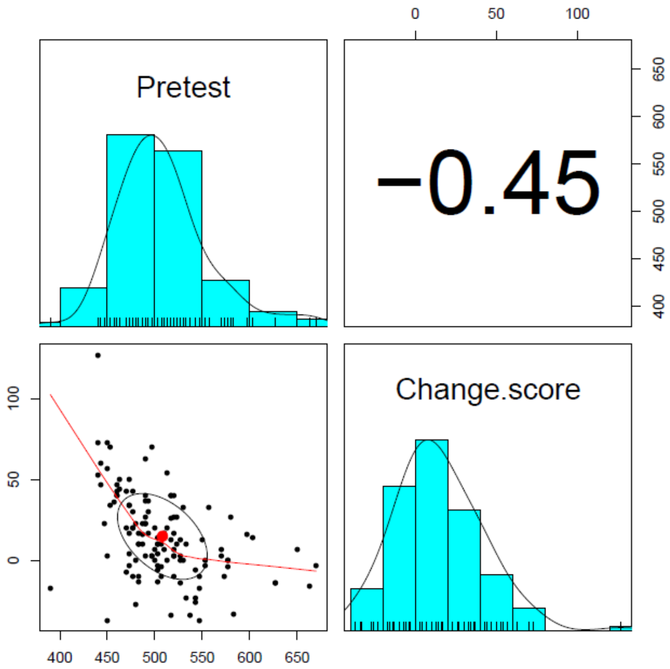

Figure 5. Scatterplot and correlations in 2013.

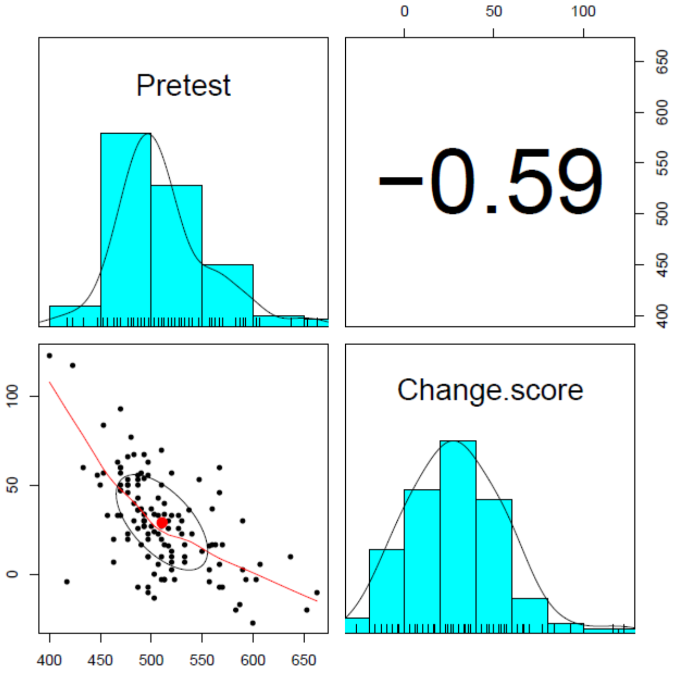

Figure 6. Scatterplot and correlations in 2014.

RQ2: What percentage of students increased or decreased their scores beyond RTM?

Table 3 compares the expected and actual posttest scores. The rightmost column presents favorable results, showing the number of students whose posttest scores exceeded the predicted ones on the basis of RTM. Across three years, approximately half of the students (51.67%, 49.60%, and 50.00%, respectively) showed a posttest score gain that exceeded that forecast by RTM.

Table 3

Comparisons of Expected and Actual Posttest Test Scores

	Expected > Actual	Expected = Actual	Expected < Actual
2012 (*n* = 120)	56 (46.67%)	2 (1.67%)	62 (51.67%)
2013 (*n* = 125)	62 (49.60%)	1 (0.80%)	62 (49.60%)
2014 (*n* = 126)	61 (48.41%)	2 (1.59%)	63 (50.00%)

Note. Expected = Expected posttest score; Actual = Actual posttest score. This note also applies to Table 4.

RQ3: What percentage of students increased or decreased their scores beyond the SED?

We used the SEDs of the 68% probability of score variability, and the values differed across years: 11.46 in 2012, 13.39 in 2013, and 12.72 in 2014. Table 4 shows that in 2012, 26.67% of the first-year students had lower actual posttest scores than their pretest score minus SED, 38.33% had actual posttest scores between their pretest score minus SED *and* their pretest score plus SED, and 35.00% had higher actual posttest scores than their pretest score plus SED. Thus, we can see in the fourth column that 35.00% of the first-year students in 2012, 47.20% in 2013, and 72.22% in 2014 had higher scores than the 68% probability SED. As mentioned by the reviewer, we are statistically supposed to have 16% of the students above the 68% probability SED (as 32% should be outside the 68% probability range and 16% should be above this range). The percentages 35.00–72.22% were all beyond 16%, which can be interpreted as a favorable result in examining growth.

Table 4

Comparisons of SED Range and Actual Posttest Test Scores

	Actual < [IPS – SED]	[IPS - SED] ≦ Actual ≦ [IPS + SED]	[IPS + SED] < Actual
2012	32 (26.67%[a])	46 (38.33%)	42 (35.00%)
2013	15 (12.00%)	51 (40.80%)	59 (47.20%)
2014	5 (3.97%)	30 (23.81%)	91 (72.22%)

Note. IPS = Individual pretest score. [a]32/120*100.

Further, we combined the two results based on RTM and the SED and investigated what percentage of students increased their scores beyond both RTM and the SED. Although not reported in the tables, we found that in the case of 68% probability, 33.33% (*n* = 40) of the first-year students in 2012, 40.00% (*n* = 50) in 2013, and 46.03% (*n* = 58) in 2014 had higher scores than those forecast by RTM and the SED. Therefore, from the viewpoint of RTM and SED, more than one-third of the students earned scores that exceeded both RTM and the SED. Thus, we can reasonably claim that such students indeed increased their ability.

Conclusion

The current study aimed to examine students' genuine growth in English-language proficiency, while considering regression to the mean (RTM) and the standard error of difference (SED). It examined RTM by (a) using correlations between change scores and pretest scores and (b) calculating the percentages of students whose posttest scores exceeded those predicted by RTM. It also examined the SED by (c) calculating the percentages of students whose posttest scores exceeded the SED range. We found moderate degrees of RTM, but also found that more than one-third (33.33–46.03%) of students increased their scores beyond RTM and the SED. Additionally, we discussed the importance of considering RTM and the SED in analyzing pretest–posttest data.

In response to the relative dearth of studies that address these two artifacts, this study has shown how they can be examined, by using real-world data. The equations we used are simple, and the necessary values can be estimated by using simple statistics; the results were useful in providing stronger evidence of claims of growth. In practice, this in turn allows teachers and researchers to offer feedback to students with greater confidence. This is particularly true of the methods we used to investigate RQ2 and RQ3, in which each student's change in scores was examined against the change expected from RTM and SED.

Our study has three limitations. First, to use Equations 1 and 2 to calculate SED, we used the reliability of the TOEFL ITP, as reported by Educational Testing Service (n.d.). That report does not provide details of the examinees from whom the reliability had been calculated (e.g., the number of examinees or nationalities); however, we assume that the reliability of .96 was higher than that we would have obtained had we had access to the raw data, because reliability estimates publicly reported are usually based on representative samples of the populations of test takers. If we had used a reliability lower than .96, the SEDs would have increased. Thus, our current result using the smaller SEDs may have made more students seem to have improved more than they actually did (see also Notes 3 and 4). Second, although we had had only pretest–posttest scores of the TOEFL ITP, access to other instruments that relate to RTM could have allowed us to better examine the impact of RTM on change in pretest–posttest scores. For example, if there were covariates that were assessed during the period when pretest scores were assessed, we could have used additional statistical analyses that can take RTM effects into account. For example, analysis of covariance (ANCOVA)—a combination of analysis of variance (ANOVA) and regression analysis—allows researchers to adjust posttest scores that are affected by RTM by utilizing as covariates data other than pretest–posttest scores (e.g., Bonate, 2000; Chuang-Stein & Tong, 1997). Further, structural equation modeling (SEM) allows one to control for measurement error when constructing a model, and this can help researchers interpret posttest scores more precisely (e.g., Kline, 2011). Third, we used a single-group, pre–post design with no control group; this design is not desirable, given a number of threats that undermine causal interpretation (Campbell & Kenny, 1999). Replication studies that feature a control group are needed, if we are to more rigorously examine proficiency gain.

We hope that our examination of two types of statistical artifact is helpful for teachers and researchers who are interested in examining proficiency growth.

Notes

[1] Instead of using SED, we could instead use two SEMs from two tests by determining whether the two confidence intervals of the two SEM ranges overlap. We did not use this method, as the SED seems to be more widely used. Readers interested in the SEM method should refer to note 2, below.

[2] The Equation differs when calculating the SED for 68% probability in comparing *different* test-takers' scores on the same test as follows:

$$\sqrt{2} * [SD \text{ of the test}] * \sqrt{(1 - [\text{Reliability of the test}])} = \sqrt{2} * \text{SEM}$$

Interested readers are directed to Harvill (1991) for details.

[3] As pointed out by the reviewer, the SEDs used in this study (11.46 to 13.39) may be a lower estimation. Educational Testing Service (n.d.) states that the SEM of the TOEFL ITP is 13.0. The SED is usually larger than the SEM, because the SED considers errors (SEMs) of a pretest and posttest (typically shown in the Equation in Note 2; however, this is an equation for a different purpose). The lower SEDs values may have been caused by the use of the higher reliability of .96, which was likely derived from representative samples of the populations of test takers with a larger range of proficiency levels than ours that were based on students at one university. Although Educational Testing Service (n.d.) does not report the type of test takers used to compute this value, we assume that the reliability was higher as values publicly reported are usually based on representative samples. However, this is the only value we could obtain because we require raw data to calculate the reliability based on the students in this study. Because we used a possibly inflated value of reliability in the calculation, our SEDs may be lower. Thus, this is one of the limitations in our study.

[4] It should be noted that the Equation in Note 2 cannot be used in our context because it should be used in comparing *different* test-takers' scores on the same test. On the other hand, Equations 1 and 2 should be used in comparing two scores from the *same* test taker (Harvill, 1991).

Acknowledgments

This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C), Grant Number 26370737. We are deeply indebted to the editor, the reviewer, and Atsushi Mizumoto for his invaluable comments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bonate, P. L. (2000). *Analysis of pretest–posttest designs*. Boca Raton, FL: Chapman & Hall/CRC.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford, UK: Oxford University Press.

Chuang-Stein, C., & Tong, D. M. (1997). The impact and implication of regression to the mean on the design and analysis of medical investigations. *Statistical Methods in Medical Research*, *6*, 115–128. doi:10.1177/096228029700600203

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—OR should we? *Psychological Bulletin, 74,* 68–80. doi:http://dx.doi.org/10.1037/h0029382

Educational Testing Service. (2007). *TOEIC user guide—Listening and reading.* Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

Educational Testing Service. (2008). *TOEIC examinee handbook—Listening and reading.* Ewing, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

Educational Testing Service. (n.d.). TOEFL® ITP. Retrieved from https://www.ets.org/s/toefl_itp/pdf/toefl_itp_score.pdf

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage.

Fulcher, G. (2010). *Practical language testing.* London, UK: Hodder Education.

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement [Instructional topics in educational measurement]. *Educational Measurement: Issues and Practice, 10*(2), 181–189. Retrieved from http://www.ncme.org/pubs/items/16.pdf

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education, 38,* 583–616. doi:10.1080/03054985.2012.731208

Marsh, H. W., & Hau, K.-T. (2002). Multilevel modeling of longitudinal growth and change: Substantive effects or regression toward the mean artifacts. *Multivariate Behavioral Research, 37,* 245–282. doi:10.1207/S15327906MBR3702_04

Mizumoto, A. (n.d.). *Langtest.* Retrieved from http://langtest.jp/#app

Ogasawara, O. (2014). *R graphical manual.* Retrived from http://rgm3.lab.nig.ac.jp/RGM/R_rdfile?f=psych/man/pairs.panels.Rd&d=R_CC

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning, 64,* 878–912. doi:10.1111/lang.12079

Revelle, W. (2014). An overview of the psych package. Retrieved from http://cran.r-project.org/web/packages/psych/vignettes/overview.pdf

Roberts, A. O. H. (1980). Regression toward the mean and the regression-effect bias. In G. Echternacht (Guest Ed.), *New directions for testing and measurement* (Vol. 8, pp. 59–82). San Francisco, CA: Jossey-Bass Inc.

Rocconi, L. M., & Ethington, C. A. (2009). Assessing longitudinal change: Adjustment for regression to the mean effects. *Research in Higher Education, 50,* 368–376. doi:10.1007/s11162-009-9119-x

Rogosa, D. (1995). Myths and methods: “Myths about longitudinal research” plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Swinton, S. S. (1983). *A manual for assessing language growth in instructional settings*. TOEFL Research Report, 83–17. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/research/policy_research_reports/rr-83-17_toefl-rr-14

Classical test theory or Rasch: A personal account from a novice user

Jim Smiley
sendaismiley@gmail.com
Tohoku Bunka Gakuen University

Abstract

Educators have utilized Classical Test Theory (CTT) when developing instruments for measuring and assessing pedagogic data. Results derived from standard CTT analysis methods offer valuable insights into the effectiveness of language assessment tools. Tests undergo a series of steps running from initial draft production through test trialing to test revision. Such instruments produced using this method can be shown to be more internally consistent and deliver more valid results than tests that are written *ad hoc* and informed by intuitive rationale. More recently, the Rasch model has gained a following among test developers as an alternative procedure in refining testing vehicles. CTT contrasts with the Rasch model in a number of key areas, differences that, when utilized in the analysis of a test, result in the production of a more internally valid test. This article questions the need for a materials developer to change to Rasch given that the learning curve is steep considering the additional investment of the time necessary to become proficient using Rasch. The conclusion is that Rasch data provides very detailed information that is *sine qua non* for long-term test instrument refinement and materials development, and that CTT data may be enough to begin the test of the test.

Keywords: Rasch modelling, Classical Test Theory, comparisons, test analysis

In this article, I document my ultimately rewarding and educational experience of attempting to broach the topic of Rasch modelling. The learning curve was extremely steep, and even after months of grappling with the core concepts, I can only lay claim to have scratched the surface and to have understood only the merest sampling that the fuller knowledge of Rasch offers. My story may be of interest to those TEVAL members who haven't yet given the Rasch model a go. However for the majority of readers, please let these words act as a guide to the frustration, the angst, the terror that many of your numerical literacy challenged colleagues feel when faced with a bewildering array of figures, of equations, of being asked to grapple with numbers.

I found the introductions to Rasch maddening in their assumptions about the readership of these primers. To qualify this statement, I need to describe my background and then tease out some of the gaps between what was expected of me and what the introductory guides expected. For many years, I have relished opening Microsoft Excel after the exam sessions to collect basic descriptive statistics. After that, I use R to create boxplots, histograms and other visuals to share information among the other language teachers. Usually, I do simple item difficulty and item discrimination analysis to find possible issues in the test construction, to highlight items that are problematic for various reasons. My Excel template has rows for student data, for total correct, for percentage correct on each question option, and so on. I can see at a glance, for example, that Question 7 (a four-option multiple choice question) was answered correctly by 67% of students, 21% choosing option A, 8% choosing B, and 4% choosing D. Both Excel and R give me (albeit with slightly different definitions and therefore results) things like the mean, standard deviation, the quartiles and so on. Books like Brown's (2005) *Testing in Language Programs* and the older tutorial book *Testing for Language Teachers* by Hughes (1989) don't faze me at all. I devoured those texts.

For more complex statistics, I use R. After making sure the conditions are met for the various tests, I generate *p*-values for *t*-tests, regression models, chi-squares and so on. I can't describe without referencing a statistics book, for example, what the conditions for ANOVA are, or the exact cut-off value that means I need to use non-parametric tests. And to learn the underlying equations for these tests would require that I study mathematics that I haven't touched for over 30 years! But I had to do that when I opened the first chapter of Bond and Fox's *Applying the Rasch Model* (2007). Without defining key terms, they render their book opaque to the uninitiated. Holster and Lake (2014) presented the basics in a much more readable form, yet by the second page, terms such as *over-fit*, *stochastic*, and *deterministic data* appear. The assumption behind these inclusions must be that the general meanings of the terms outside testing cover the specialist meaning sufficiently. For this reader, I'm afraid that they don't. *Shiken* has a set of resources aimed at introducing Rasch measurement to members (Sick, 2008a, 2008b, 2009, 2010). Once again, the opening pages read as text written for insiders. The manual that came with the Winsteps software needs at least high-school algebra to comprehend. I appreciate the fact that there are concepts, techniques and methods to be learnt. But perhaps there is a cultural gap also at play here, and the in-crowd either not realizing there is or not wanting to overcome the cultural divide.

The version of this article is the result of revisions after two anonymous reviewers commented on an earlier draft. I thank them from the bottom of my heart for their efforts. Both provided copious notes, suggestions for improvements, pointed out errors in my conceptualization of Rasch principles, and generally added significantly to my understanding of Rasch. Reviewers such as them add to the joys of learning. However, this article is still bound to amuse Rasch purists, who will certainly find many misunderstandings remaining. I would highly appreciate those errors to be pointed out and corrected in a later issue of *Shiken*. Perhaps if more novices were to share their stories, TEVAL may become more of a beginner-friendly SIG.

Total test scores from Traditional or Classical Test Theory (CTT) have been described as "simple raw scores" (Holster & Lake, 2014). A test-test taker's final score is obtained through the addition of their correct raw responses. This total and the total of all other test takers in a test session combine to produce data which the test developer uses for analysis. These "group-centred" scores form the basis for statistical analysis and "require the clustering of individuals into discrete categories or populations" (Choppin, 1983). The focus on the group allows for statistics that rely on the nature of that group, not on the specifications of the test instrument itself. For example, using data from a population or a sample, one can easily discover the interquartile ranges, the variance of the mean, whether or not the samples' means are statistically significantly similar or different and so on. With a different sample set, the figures returned deliver another set of statistics. These data provide the test developer with some tools to analyze the validity of the test, but they do not allow for a complete understanding of the validity.

This lack of interface between the test instrument and the resultant raw scores is problematic for test developers. Students are measured on the basis of what may have been a faulty test, yet the absence of technical analytic tools hinders the discovery of a potentially flawed test. Flaws also include reliability issues such as the test actually measuring what it tried to measure (construct validity), and the test question types targeting the skill appropriately (face validity) (Hughes, 1989, pp. 26-27), but a discussion of these is outside the scope of this paper.

Furthermore, test developers need to be able to test their tests independently of ability of the test takers. A stable test returns similar results irrespective of the particular group of students. It behooves the developer to ascertain the reliability of the test and to ensure that the test is able to perform its function. A poor test may be testing irrelevant content, or the manner of the writing may be uncritically biased towards a particular ability level for reasons that are not related to the test but to the quality of the writing. In such cases, the test instrument loses some of its usefulness. A non-test example of an inappropriate instrument would be a measuring jug made of paper used to measure the volume of boiled water. The test instrument, the paper jug, is an inadequate vehicle for its purported task.

CTT theorists have developed methods to overcome these barriers (Brown, 2005). The twin tools of Item Facility (IF) and Item Discrimination (ID) attempt to go beyond the nature of the total score and investigate more detailed relationships between the individual items on a test and the overall scores. IF and ID are relatively easy to understand even for those without a background in statistics. They can be obtained using spreadsheet software, such as Microsoft Excel, with only a minimum amount of preparation when all the raw data is collated. Split-half reliability analysis helps test writers understand the balance of a particular test, where the difficult items are found in a single test. The Rasch model is predicated on the individual at both the level of the test item and the test taker. Various software tools exist that allow detailed analysis of raw data according to the Rasch model. Winsteps (Linacre, 2014) was used here. Using the tables, diagnostic tools, graphs and other functionality available in Winsteps requires at least a solid command of basic statistics and measuring methodology. Its learning curve is steep.

This article attempts to answer the question: is the information provided by Rasch significantly more valuable than CTT given the time required for its study? In other words, is Rasch's payoff enough to justify the time spent? A case study is shown in which a test is subjected to CTT analysis and Rasch analysis. The types of information arising from each analysis are discussed, and the pragmatic decision about the use of CTT and Rasch is given.

CTT provides tools that analyze overall test scores and that aim to judge the whole test holistically. Item facility describes the easiness of any individual test item, item discrimination shows how well an item did in separating the high scorers from the low scorers, and split-half reliability expresses the degree to which subsets of items provide consistent ranking of person ability. Following Brown (2005, p. 66), to calculate item facility (IF) for each item, the total score obtained by each student is divided by the total number of students.

$$IF = \frac{Total\ Correct}{Total\ Number\ of\ Takers}$$

If all test takers got the item correct, IF = 1.0. Correspondingly, if all test takers were mistaken, IF = 0.00. This simple tool can highlight test items that were too difficult or too easy. In Excel, test data can be sorted by IF score. Then the relative number of easy-to-difficult items can be ascertained. Using this, the balance of item difficulty, or facility, can be understood.

Item Discrimination (ID) develops on IF (Brown, 2005, pp. 68-70). A percentage of the examinees is chosen, usually between 25% and 33%. The IF scores of those test takers who scored in the bottom 25% (or 33%) is subtracted from the IF scores of the top 25%.

$$ID = IF(Top\ 25\%) - IF\ (Bottom\ 25\%)$$

Top scorers in a testing group should score higher than low scorers. Test items that distinguish well between these two groups, i.e. when ID < .4 (Brown, 2005, p. 75, citing Ebel, 1979) are stable. If, however, ID < 0:0, lower scorers got the item right more often than higher scorers. When this happens, the item needs to be analyzed to see why this happened.

Split-half reliability (SH) provides an estimate on the overall test reliability (Brown, 2005; Hughes, 1989). Test reliability is a function of both halves of the test resulting in equal scores for each student. On a 100-item test, any individual student can be given two scores:[1] Score 1 comprising the total correct from the odd- numbered questions, and Score 2 comprising the score from the even-numbered questions. If the test is reliable, Score 1 should be similar to Score 2 (Hughes, 1989, p. 32). For example, Student 1 scores

[1] This SH method is the second Hughes (1989) describes and is more robust because his first method of generating score 1 from the first 50 items and score 2 from the latter 50 is problematic for tests whose questions get progressively more difficult deliberately.

38 on the odd-numbered items and 36 on the even-numbered items on a 100-item test. There is not so much discrepancy between these two halves. There is, however, an inconsistency in the results in the test when Student 2 scores 38 on the odd-numbered items and 16 on the even-numbered items.

Split-half reliability speaks more to overall test imbalance than to item or person analysis. Its use as a test of the test is vindicated in that it can point out imbalances in the test design. Brown (2005) also suggested the Cronbach alpha coefficient as another way to calculate reliability, but cautions that "conceptually, the split-half method is the easiest of the internal-consistency procedures to understand" (p. 179).

Winsteps (Linacre, 2014) offers a wide variety of functionality for many different levels of analysis. This section describes five key tools that offer the most immediate benefit to test developers and are the most accessible in that they do not require knowledge of advanced statistics. In a similar way, most users of a software tool such as Adobe Photoshop only use a small subset of that program's functionality. Georg Rasch wanted a method that understood the role and position of the individual within the frame- work of the construct under investigation.[2] "Individual-centred statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated." (Rasch, 1960, p. vii, cited in Choppin, 1983, p.12)

The Rasch model is described mathematically by an equation that predicts the expected score of any individual on any item on the test based on a matrix of item responses from all candidates. This necessitates a complex calculation, and Winsteps performs multiple iterations before settling on the model. The steps below summarize roughly what Winsteps does to calculate data-model fit. These are shown linearly to suit the medium of an article. In real terms, though, the program performs the steps many times, through a number of iterations until it has reached an acceptably low Maximum Score Residual (MSR).

1. Raw score data is read.
2. Each test taker's total score is tallied.
3. Each item's item difficulty level (or item measure) is calculated.
4. Items are placed on a scale of difficulty.
5. Each individual's item-by-item expected score is worked out (i.e. their overall score places them at a particular point on the scale, and this is judged against the difficulty of each item).
6. The score residual is calculated. This shows the difference between the total of the expected scores against the actual scores.
7. The model is refined through a process of updating the subsequent iteration using the information derived from the current one.
8. The iterations stop when the MSR reaches a pre-set level.

Any mathematical model must make assumptions. There is a critical difference in the assumptions underlying IF in CTT and how Rasch works out the difficulty of an item. We have seen earlier how IF is derived in CTT, as a function of the total correct answers divided by the total number of questions. No individual's overall score is factored in. Because of this, CTT requires a further step: the calculation of Item Discrimination (ID). With ID, the test analyst must decide on a percentage of high and low scorers. Using the IF for each group, ID can be established. ID figures are highly useful in discovering faulty test items, but even with IF and ID, it is pragmatically difficult to judge whether an individual test taker got a question right or wrong by luck or by guess.

[2] It is beyond the scope of this article to discuss the mathematical formulae that describe the Rasch model nor develop a discussion into the history and principles behind Rasch's statement. For further reading into the development of the Rasch model, see Bond and Fox (2007) chapters 2 and 3.

A critical difference between CTT and Rasch is that Rasch accounts for the fact that some test takers' responses on test items do not reflect their true ability. An examinee may guess correctly on an item that they have no idea about. Alternately, a high scoring test taker may slip on a relatively easy item and get the item wrong. The concept of the "expected score", then, is crucial to understanding how Rasch decides on the probability of a response type per examinee per item. The sum total of an examinee's correct responses shows that participant's general level. The difficulty of an item is measured against the level of the examinee, and the probability of that person getting that item correct can be understood. In other words, the item difficulties can be used to estimate the probability of any person answering any item correctly.

Two variables are involved, the test taker and the test questions. Let's look at each variable in turn using a test of 10 questions. The test taker can be in one of three states: they are too good for the test, in which case they will score 10/10; they are too bad for the test, scoring 0/10; or they are somewhere in the middle. Both Rasch and CTT effectively ignore those who are too good and those who are too bad. In CTT, test takers are given scores of 100% and 0% respectively. In a sense, CTT does ignore these scores as 0% and 100% are not meaningful beyond a purely ranking measure. Rasch labels these test takers as "extreme" and their data does not contribute to measurement. Another way of expressing these extreme cases is to say that the test does not adequately measure their level. More difficult test items are needed for the high ability examinee, and more simple items needed for the other. The information about the test given by CTT and Rasch analyses can only be effectively utilized when test-test taker scores fall within 1 to 9.

An examinee scored 9/10. They made a mistake on one test item, but overall their final score is as high as possible that is useful for analysis. We need to wonder about the item that was wrong. Did they not know the information in the question, or did they slip up? A test taker who scored 1/10, similarly, may have known the question genuinely or simply guessed. Yet a test taker who scored 9/10 is of a higher level than one who scored 1/10. There may be times in a test when a test taker guessed an answer correctly and other times when they slipped up on a question that is easier than their level. These responses are said to be "unexpected". In order to judge this, we must be able to analyze item facility (IF). The model may be summed up thus: the probability of any student getting any question correct is a result of the difference between item difficulty and person ability. A feature of the Rasch model include is a test taker's total correct score provides rank ordering of ability. In other words, a score of 9/10 indicates higher ability than 1/10, even if there are questions that were answered unexpectedly.

I have selected three tools to demonstrate some of the functionality of Rasch. I believe all three of them to be conceptually simple; they may all be understood without an advanced knowledge of either the underlying mathematical model and they produce values that appear ranked and may be understood as so without losing too much of the inherent subtlety. I have used these to show the power of Rasch quickly and successfully to colleagues far more mathematically challenged than myself.

Point-measure correlation is in some ways similar to ID in CTT. Point-measure correlation refers to the correlation between the difficulty of each individual item and the difficulty of the test as a whole. A value of 1.0 would indicate that all low ability test takers got the item wrong and all high ability test takers got it right, that is it indicates a perfect correlation between the item responses and the estimated Rasch measures of the test takers. A value of zero tells developers that there is no relationship between the particular item's responses and the rest of the test. In other words, whether students got it right or wrong is random. A negative value indicates a flawed test item as the lower scorers got that item correct more often than high scorers. These negative values are more problematic than zero values, and may indicate that the item is flawed in some fundamental way, and that it should be checked to see whether the answer key was wrong, revised, or possibly deleted from the test.

A subtlety that may be missed by novice Rasch users is its apparent ranking method. When classroom teachers see percentage scores, they may interpret them as representing equal intervals on a line from 0 to 100. Yet CTT does not attempt to show the interval between, say, 45% and 46%. Depending on the test, the interval between these two scores may well be virtually nothing, or it may be very wide. Rasch, on the other hand, aims to provide equal interval measures, so Rasch point-measure correlations are based on interval level measures, whereas CTT ID values are not. Teachers may miss this subtlety, but the concept of more difficult and easier items is not challenging.

Table 1 shows typical Winsteps item statistics. Reading from the left, we have the item number, the total score (the number of correct responses), the count of all responses, and the logit measure of item difficulty. No examinee could answer #10 accurately and the estimated measure of 101.87 is thus an extreme score. As mentioned earlier, there is a conceptual gap between these measures which look like percentage figures and the real workings of Rasch. Part of this apparent similarity can be explained by Holster and Lake (2014, p. 140) who suggested setting the mean item difficulty at 50.00 because "figures in the range of 50 to 100 are easier to understand", whereas according to them, researchers "usually set it to 0". But even if these measures are not percentage values, they generally fall within what looks like figures non-Rasch specialist classroom teachers are likely to comprehend. This table is ordered from the most difficult item first then successively adding the easier items. Other orderings are possible (for example, tables ordered by the closest match of the items' measures to the model). The measure values give a ready understandable account of the relative difficulty of each item. The total score figures rise as the measure value falls.

Table 1
Rasch Item Statistics

				Model	Infit		Outfit		Pt measure		Exact Match	
Item	Score	Count	Logits	S.E.	Mnsq	Zstd	Mnsq	Zstd	Corr.	Exp.	Obs%	Exp%
10	0	16	101.87	18.75		MAXIMUM	MEASURE		.00	.00	100	100
8	2	16	77.81	8.67	0.54	-0.8	0.25	-0.4	.68	.46	93.8	89.6
13	3	16	71.33	7.49	0.82	-0.3	0.63	-0.1	.58	.50	87.5	84.8
14	3	16	71.33	7.49	1.04	0.2	1.01	0.3	.46	.50	87.5	84.8
12	4	16	66.24	6.84	1.37	1.1	2.01	1.3	.27	.53	75.0	80.4
15	4	16	66.24	6.84	0.58	-1.3	0.36	-1.0	.76	.53	87.5	80.4
11	7	16	54.23	6.07	1.36	1.2	2.45	2.4	.27	.55	68.8	75.5
1	9	16	47.12	5.99	0.95	-0.1	1.15	0.4	.52	.52	81.3	74.2

Rasch variable maps, or Wright maps, such as shown in Figure 1, plot the test taker and item on a vertical scale according to the test taker's ability and the item's difficulty. High scoring test takers and difficult questions are at the top. Using the visual data, the test developer can a number of kinds of information. Because the data is visual, there is an immediacy to its interpretation that novice users and classroom teachers appreciate. Items that have no corresponding test takers at the top are too difficult and are not useful in segregating populations of higher ability test takers. A few items that are above the level of the examinee group are needed to ensure no ceiling effect. Those items at the bottom are too easy and offer no useful information about the level of the lowest ability test takers. Too many test takers lined up with a single question points to the lack of questions available to discriminate between those test takers. Too many questions for too few test takers indicate that there are too many questions at the same level, again an indication that the test items need to be analyzed for purpose.

Figure 1 uses the same data set as Table 1. Visually, it can be seen that Item 10 is right at the top of the map, and the same downward ordering of the questions' difficulty is observable on the right-hand side. Here we also have student data. As well as the measure of the question item, Rasch also computes a measure for each test taker. These values are positioned on the left-hand side of the map. S14 is the highest ability examinee and S01 the lowest.

```
MEASURE      Person - MAP - Question
                 <more>|<rare>
   90                  +  Q10
                  S14  |T
   80                 T+
                       |  Q8
                       |
   70             S12  +S Q13    Q14
                      S|  Q12    Q15
   60   S03  S06  S15  +
                  S10  |
             S02  S07 M|  Q11
   50                  +M
             S04  S13  |  Q1     Q4
   40             S09  +
        S05  S08  S11 S|  Q3
                       |  Q5     Q9
   30             S01  +S
                       |  Q6
   20                 T+
                       |  Q2
                       |T
   10                  +  Q7
                 <less>|<frequent>
```

Figure 1. Person-item map showing student ability on the left and item difficulty on the right.

The table of distractor frequencies, shown in Table 2, is arguably the single most useful tool which can be interpreted without too much conceptual difficulty for the novice Rasch user. This table shows the number of test takers that selected each particular option for every question. Also, the average ability of test takers for each option is shown. Together, these provide a highly useful tool for the refinement of a test vehicle. Misfitting items are marked with an asterisk. These are items where the correct option was selected more by lower ability level (on average) than not. I use 4-option multiple choice items as an example in Table 2. In the first part of the table, Winsteps shows the item number, the options (here A = 1, B = 2 and so on), and a score value, which is the correct answer. The 1 is always at the bottom of the set. Next to these values next is the data count in both raw figures and percentages of the total test takers. Item 10 was answered correctly by no examinee. Option A (i.e. data code 1) was selected by students whose measured ability averaged 55.04. Option B by students at 47.34, and Option C by examinees at 51.23. The spread of the selection is reasonable. No single distractor monopolized the selection. This can be contrasted by looking at Option B in item 8. Only one examinee chose that and their measured ability level was low.

Item 14 highlights a potential problem in the test. The ability of those examinees who answered correctly as 65.62. Yet a higher ability test taker (at 69.73) chose another answer. The asterisk provides an immediate clue to this problem. In this case, only one higher level test taker made an error, and it is likely that this was simply a slip. But, if there were many higher ability examinees choosing the wrong answer, that is a serious indication that the item needs to be investigated.

Table 2

Distractor Frequencies

ITEM NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTFIT MNSQ	PTMA CORR.
Q10	2	0	4	25	47.34	5.91		-.16
	3	0	7	44	51.23	5.36		-.01
	1	0	5	31	55.04	9.28		.16
Q8	2	0	1	6	36.46		0.10	-.26
	1	0	7	44	48.08	4.35	0.60	-.20
	4	0	6	38	49.09	5.27	0.70	-.12
	3	1	2	13	77.79	8.06	0.20	.68
Q13	2	0	3	19	43.68	7.23	0.50	-.25
	3	0	4	25	47.33	6.68	0.70	-.16
	1	0	6	38	49.09	5.27	0.80	-.12
	4	1	3	19	69.41	9.59	0.60	.58
Q14	4	0	8	50	43.92	3.56	0.50	-.51
	1	0	4	25	51.29	7.26	1.00	-.01
	2	0	1	6	69.73		3.60	.32
	3	1	3	19	65.62*	11.19	1.10	.46

Method

Participants, Materials, and Procedure

Case study data are taken from a test written to supplement the author's English language textbook *Nursing Care* (Smiley & Masui, 2013). This textbook is designed for students on a nursing course at the university level studying English as a part of their curriculum. The prior English language level assumed at the start of the course is roughly between Grade 3 and Pre-Grade 2 *Eiken*. The *Monkagakusho,* the Japanese Ministry of Education, states that the target finishing level of high-school pupils should be *Eiken* Pre-Grade 2 (MEXT 2013), so this book is considered suitable for the university English course. The test, comprising 50 multiple-choice items, assesses Units 1 to 6 of the book, and there is a further Test B for units 7 to 12. This case study looks only at the first test. They are considered criterion referenced tests (CRT) (Hughes, 1989) as students have finished the relevant units before taking each test. However, there is a degree of norm-referenced type material present. Students at university exhibit a large range in English proficiency, so a published textbook for this level contains material many students have not yet mastered. Ideally, a CRT only assesses elements that were new to students at the start of the course, but in this case because many students did not have a Grade 3 ability prior to the start of the course, a significant amount of the erstwhile assumed language and the technical nursing content were new.

Results

CTT Results

As shown in Table 3, the test produced an average score in the 50% to 60% range. CRTs may be expected to return higher scores if the language is known prior to the start of the course and the test vehicle assesses only the new content. As mentioned above, however, there is a sizeable number of students who have not attained a proficiency level of *Eiken* Pre-2nd Grade. Their task throughout the course will be to simultaneously learn the test content and develop their basic language proficiency. With this taken into consideration, an average of 52.2% may be considered acceptable.

Table 4 shows those items that have the top five and the bottom five IF scores. Items 24 and 15 are above .85 which indicates that they are easy. Item 37 was only answered correctly by 8% of test takers. This item needs to be investigated. Items 29 to 26 all return a score under .2, and they may also be too difficult. IF shows the test developer that there are certainly three items that require thought and perhaps alteration and five or six others that need further analysis before their place in the test is assured.

Table 3
Nursing Care Test Summary Statistics

	mean	SD	Max	Min
Score (Max.50)	26.1	6.8	41	12

Table 4
Item Facility Values

	Top 5					Bottom 5				
Item number	24	15	18	40	9	26	47	44	29	37
Item facility	.87	.86	.8	.8	.79	.19	.17	.14	.1	.08

Theoretically, every student has access to all the information that will be in a criterion referenced test during the course duration. Learning objectives are specified prior to the teaching term and learning actions are chosen to allow for maximal retention of those objectives. A textbook is a set of learning objectives that contain learning actions. Therefore a test that is wholly based on a textbook must be defined as a criterion referenced test. Higher IF scores may be expected than from a norm-referenced test where the items may be drawn from language or content elements examinees have yet to encounter. Conversely, very low IF scores may be indicative of a number of serious issues in the class: students may be unmotivated to learn the material in the textbook, the assumed starting level of the student body may be too high, the class content may have focused on segments of the book that were not targeted in the test. The test items themselves may be too obscure, in that they test too narrow areas of the book, topics or language that appears only once.

Question 29 highlights another test writing difficulty. Only 10% of takers got this item correct. The question's distractors A, B, and C, are all possible answers. The correct answer (D) reads, "All of the above". Examinees were not accustomed to this question type, and it only comes once in the test, so they could not train themselves to expect this type. Upon investigation, Question 37 throws up another issue in test validity. Many tests use the form:

Q37: It is important to be __________ to new patients.
a) helpful b) helping c) helped d) helper

This kind of item seems intuitively useful to many teachers. All verb forms and the noun form "helper" may be in the category of assumed knowledge. Yet, something inhibited examinees from answering correctly. Anecdotally, because the subject matter is sensitive in our institution, I can report a heightened discussion over the use of this type of item when a post-test study revealed that a similar IF score was returned in our entrance exam. Perhaps the control examinees have over verb conjugation is not strong enough to merit a test item that focusses only on that. Discrete point testing may be less valid as a measure of holistic ability than is believed at my institution. Question 44, similarly, offers a counter-intuitive response, this time on the discrete testing of a noun item.

Q44: Aerobics is a good way of keeping __________.
a) exercise b) fit c) health d) lifestyle

IF is seriously limited in its ability to show the distractor selection ratio. That 14% of test takers chose B is known. The ratio chosen for the others is necessary before any assessment can be made.

ID values are summarized in Table 5. Brown (2005) provided guidelines on item discrimination as to which items do a good job in discriminating between the high and the low scores. An ID score of .40 and above indicates a solid item. Scores of between .39 and .30 are considered good. Items whose scores are between .29 and .20 need some alteration. This change depends on whether the item should be made more difficult or easier. The judgement for this action is based on the numbers of test takers scoring accurately in each high or low group. The ID score itself does not give information directly; the writer needs to look at the precise details of the responses for that item. Scores below 0.20 do little to differentiate between the higher and the lower groups. In this test, one item (Question 27) had a negative correlation score. This means that the lower group students scored more highly than the higher group. This item needs to be changed.

> *Q27: Why did Sara stand on some scales?*
> *a) to let the nurse measure her weight*
> *b) to let the nurse measure her body height*
> *c) to measure her weight*
> *d) to measure her body height*

Table 5
Item Discrimination

Discrimination	>.40	>.30	>.20	<.20	<.00
No. of Items	24	8	6	12	1

This question is one of three that follow a short paragraph-length reading. Even without the accompanying text, proficient users of English will be able to eliminate distractors B and D. The answer comes down to the distinction between the passive "having her weight measured" or the active "measuring her (own) weight". The text reads ". . . and the nurse measured her weight". Is this distinction too fine to be useful at this level, or is there something about the distractors that added some complexity to the question. Again, CTT does not offer ready tools to find this out.[3]

Split-half reliability is shown in Table 6. Items were split in two ways: between the first half of the test, assessing listening, and the second half, assessing reading, and between odd-numbered and even-numbered items. Both analyses returned a correlation coefficient of .68, indicating modest reliability. The average scores show that the listening section was statistically significantly easier than the reading section. A paired-sample *t*-test returned values of t = 14:45, df = 142, p < .01. The odd-even split half figures show a slightly less extreme imbalance, and the total scores are reversed.

Table 6
Split-half Reliability

	Mean Score	Correlation
Items 1-25	59.7%	.68
Items 26-50	42.1%	
Odd items	48.1%	.68
Even items	56.1%	

At many points in the analysis, using CTT tools generated more questions than answers. IF information did highlight those areas of ease and difficulty, but without ready access to the distractor selection ratios,

[3] In my pre-Rasch Excel days, I often generated this information in Excel using COUNTIF(cellrange=1), COUNTIF(cellrange=2), and so on. But manually preparing these sheets was time-consuming.

further analysis must necessarily be limited. ID is a useful tool to check if the test items inadvertently contain biases towards lower ability level test takers. Those items that fail the ID test can be analyzed further, but the same limitation applies here as to IF. Split-half reliability talks about the test as a whole, so offers very little to help the writer revise the test.

Rasch Results

Winsteps' summary statistics provide the same basic figures as can be output by Excel; the mean, Standard Deviation, Maximum and Minimum raw scores. Winsteps' Rasch summary statistics, shown in Table 7 and Table 8, provide information about both persons and items, including the logit measures. Winsteps models the persons and items as it works out the precise relationship between these, but models do not return a perfect match with real world data, so the summary statistics indicate the degree to which the data fit the model. A novice user will not know the acceptable range of values for infit and outfit. Taking Holster and Lake (2014) as a guide, the person infit and standard deviation are acceptable at 0.99 and 0.15 respectively. The corresponding outfit values also seem to be acceptable. The item infit and outfit values are similar to that of the person's, suggesting that the model is a satisfactory match to the data. One reviewer pointed out that values +/-0.30 are acceptable revealing that the maximum item infit of 1.25 is good, but the cut-off points of 1.30 and 0.70 are reached in the maximum item infit and outfit, where 1.30 can 1.38 can be seen. These values are the result of the data not matching the model, i.e. when a lower ability student got a difficult item correct. Being summary statistics, the information speaks to the test as a whole. Also shown are the Rasch reliability of separation estimates for the test and Cronbach's alpha, analogous to the split-half reliability shown in Table 6. The Rasch person reliability and Cronbach's alpha are considerably higher than the split-half reliability because they are calculated from the entire 50 items, rather than the 25 items used to calculate split-half reliability.

Table 7

Summary Statistics for Persons

	Total			Model	Infit		Outfit	
	Score	Count	Measure	Error	Mnsq	Zstd	Mnsq	Zstd
Mean	26.1	49.8	50.94	3.30	0.99	0.0	1.02	0.1
S.D.	6.8	0.8	7.14	0.16	0.15	1.0	0.24	1.1
Max.	41.0	50.0	68.69	4.07	1.43	3.2	1.70	3.2
Min.	12.0	43.0	35.63	3.17	0.69	-2.7	0.53	-2.4

```
Real Rmse    3.39 True Sd    6.29   Separation  1.85  Person Reliability  .77
Model Rmse   3.31 True Sd    6.33   Separation  1.92  Person Reliability  .79
S.E. of Person Mean = 0.60
```

```
Notes: 143 persons,  50 items
Person raw score-to-measure correlation = 1.00
Cronbach alpha (KR-20) = .79
```

Table 8

Summary Statistics for Items

	Total			Model	Infit		Outfit	
	Score	Count	Measure	Error	Mnsq	Zstd	Mnsq	Zstd
Mean	74.5	142.5	50.00	1.99	1.00	-0.1	1.02	0.0
S.D.	30.2	0.7	11.11	0.28	0.09	1.3	0.16	1.4
Max.	125.0	143.0	76.79	3.08	1.30	4.1	1.38	3.8
Min.	12.0	141.0	29.75	1.77	0.83	-2.5	0.72	-2.3

```
Real Rmse   2.05 True Sd   10.92   Separation  5.33  Item   Reliability  .97
Model Rmse  2.01 True Sd   10.92   Separation  5.43  Item   Reliability  .97
S.E. Of Item Mean = 1.59
```

```
Notes: 50 items, 143 persons
```

Figure 2 shows the Wright map comparing persons and items. No extreme items or persons were present in this test. Questions 37and 29 are the most difficult with scaled scores of about 75. The summary statistics tell us that the max person was 68.69, and this can be seen on the map. This is analogous to the IF information delivered earlier, and a similar investigation into the possible causes of the item's difficulty may be conducted. Generating the variable map took two mouse clicks. The same cannot be said for creating the IF table. IF informs about the relative numbers of test takers getting the item correct, and the variable map gives an indication of the distance between the upper (and lower) reaches of the items and the examinees. Having a test taker overall range outside that of the items would be highly suggestive of a test that did not accommodate all of the examinees' ability levels. These two tools offer similar information, and together their power contributes more to an understanding of the test.

```
MEASURE             Person + Item
                    <more> | <rare>
   77                      +  Q37
   75                      +  Q29
   72                      +T
   70                   1  +  Q44
   67                   0  +  Q26 Q47
   65                 000 T+  Q21 Q39
   63         00000000011  +  Q33
   60               00001  +S Q31 Q35 Q46
   58         00000000011 S+  Q20 Q27 Q42
   55   00000000000000111  +
   53    0000000000111111  +  Q23 Q25 Q32 Q36 Q41
   51 0000000000000011111 M+M Q14 Q16 Q49 Q5
   48       0000000000011  +  Q1 Q12 Q30 Q4  Q8
   46      00000000111111  +  Q10 Q17 Q2  Q3  Q48 Q7
   43     000000011111111 S+  Q28 Q45
   41            00000011  +  Q19 Q22 Q34 Q38 Q43
   39                0011  +S Q11 Q50 Q6
   36               01111 T+  Q13 Q18 Q9
   34                      +  Q40
   31                      +  Q15
   29                      +  Q24
                    <less> | <frequent>
```

Figure 2.
Variable map showing the distribution of persons and items.

Figure 2 shows the expected bell-curve like histogram for both items and persons. A classroom teacher may feel satisfied with this distribution. However this fails to appreciate a main purpose of a well-designed test which is to discriminate between different ability levels of test taker, so a flatter distribution of item difficulty would suggest a better discriminatory instrument than a bell curve. With all histograms, the bucket size has an important bearing on its shape. For example, Figure 3 shows a zoomed-in view of the gap around the 55 level. Using this information, analyzing questions 27 and 42 against questions 23 and 25 may allow for more precisely targeted questions around those levels to be developed.

```
MEASURE                                         Person + Item
                                                <more> | <rare>
  57                           12  27  56  89  91 110  +  Q27  Q42
  56                                   25  64  92 100  +
  55  1  5  6  19  32  46  49  57  60  88  96 116 128  +
  54                2  36  37  74  86  95 107 120 138  +  Q23  Q25
                                                <less> | <frequent>
```

Figure 3.
Magnified variable map showing items and persons between 54 and 57 scaled points.

Why is question 23 easier than question 42? Perhaps this is impossible to answer definitively, but the process of trying is valuable.

> *There is an underlined word in each sentence. Choose the best meaning from the options.*
>
> *Q23: The doctor was worried about John's diet.*
> *a) John is trying to lose weight*
> *b) what John eats on special days*
> *c) what John eats usually*
> *d) John wants to become smaller*
>
> *Q42: General hospitals have many departments __________ are very big.*
> *a) and b) too c) though d) even*

The textbook glosses the term *diet* in Japanese, and students who remember that definition are likely to select option C. Conversely, there are no direct grammar directions in the textbook, and students have no practice of conjunctions or non-repetition of the subject after a conjunction when there is no comma. Q42 may be challenging from the perspective of a Japanese learner of English through L1 interference as subjects are typically not be repeated. Japanese is a theme-rheme structured language, and syntax such as *Hospitals have many departments too very big* is acceptable. In this interpretation, the emphasising function of *even* may be placed directly after the *departments* to provide the rheme comment on the *hospital*. Or the grammatical potential for complexity may be immaterial if the difficulty is due to the focus on the discrete item which is either known or unknown.

Items 35 and 50 also show an interesting result. Both questions test knowledge of discrete vocabulary items. *Annual* and *updates* are glossed in the textbook and are recycled throughout the unit in which they appear. Both sentences are in the active voice and contain nothing out-of-the ordinary in terms of object and adverbial clauses. Intuitively, I would have estimated *annual* to be the more challenging term especially as *update* is a commonly used word in Japanese that has a very similar semantic scope to the English. Very little separates them in terms of perceived difficulty, yet Item 50 is measured at 39 and Item 35 at 60.

> *Q50: John came to the clinic for his __________ health check up.*
> *a) by year b) year c) annually d) annual*
>
> *Q35: Nurses give patients' families __________ on their health.*
> *a) new b) updates c) tests d) conditions*

The five items with the poorest point-measure correlations are shown in Table 9. One item, Item 27 has a negative value. This is the same item that was discovered by ID and discussed above. The rest are under .10. In the whole table, only 15 items have a value of .40 or above, the cut-off figure Holster and Lake (2014) recommended as showing that an item is functioning well. These point-measure values do not indicate that the test as a whole is performing well as an instrument that differentiates between different ability levels of students. The CTT ID values pointed to 12 questionable items, but Rasch highlighted 35 items that require attention.

So far, the tools have foregrounded items that deserve further investigation. At each juncture, the information regarding the ratio of selection of the distractors was missing. As a result, the test developer can focus the attention on the where but not precisely on the how. The distractor frequencies, shown in Table 10 fill in this missing piece. Winsteps has ordered this table according to the degree to which the model predicted the responses. Those items that functioned less well are at the top as can be seen with the outfit mean-square value in the third rightmost column. To a test designer, however, there are two other columns that hold very valuable information. The data counts and the average ability columns show how many test takers selected each question option and what the overall level of those test takers is. These figures provide a means by which the developer can see exactly how well the test items discriminated the various levels of test taker. An asterisk next to a value indicates that the average level of test taker getting the item correct is lower than the average of another option chosen. Ideally, high scorers select the correct option and lower scorers select the other options. This happened nine times in this test.

Table 9
Point-measure Correlations for the Five Poorest Performing Items

	Total	Total		Model	Infit		Outfit		Point-measure	
Item	Score	Count	Measure	S.E.	Mnsq	Zstd	Mnsq	Zstd	Corr.	Exp.
27	51	142	57.39	1.84	1.30	4.1	1.37	3.8	-.09	.31
33	36	143	63.01	2.01	1.16	1.6	1.38	2.5	.02	.29
26	28	143	66.52	2.19	1.11	0.9	1.36	1.9	.05	.26
44	20	143	70.84	2.48	1.11	0.7	1.28	1.2	.05	.23
37	12	142	76.79	3.08	1.07	0.4	1.24	0.8	.06	.19
47	25	143	68.01	2.28	1.12	0.9	1.27	1.4	.06	.25

Table 10
Distractor Option Frequencies for the Five Poorest Performing Items

			Data		Average	S.E.	Outfit	Point-M
Item	Code	Score	Count	Percent	Ability	Mean	MnSq	Corr.
33 A	1	0	1	1%	38.08	N.A.	0.2	-.15
	3	0	1	1%	46.63	N.A.	0.6	-.05
	2	0	105	73%	51.00	0.67	1.1	.01
	4	1	36	25%	51.24	1.33	1.5	.02
27 B	2	0	6	4%	45.37	2.69	0.6	-.16
	3	0	82	58%	51.76	0.82	1.3	.13
	4	0	3	2%	55.20	2.30	1.5	.09
	1	1	51	36%	50.03*	0.94	1.4	-.09
	MISSING	***	1	1%	50.37	N.A.		-.01
26 C	1	0	3	2%	43.55	1.05	0.4	-.15
	4	0	14	10%	43.72	1.74	0.5	-.33
	2	0	98	69%	51.99	0.64	1.2	.22
	3	1	28	20%	51.66*	1.56	1.4	.05
44 D	1	0	7	5%	44.87	3.38	0.7	-.19
	4	0	12	8%	46.94	1.51	0.6	-.17
	3	0	104	73%	51.63	0.70	1.2	.16
	2	1	20	14%	51.84	1.43	1.3	.05
37 G	4	0	5	4%	47.50	3.72	-0.7	.09
	3	0	27	19%	49.32	1.38	0.9	-.11
	2	0	98	69%	51.38	0.73	1.1	.10
	1	1	12	8%	52.25	1.72	1.3	.06
	MISSING	***	1	1%	52.76	0.02		

Looking at Question 27 again:

> *Q27: Why did Sara stand on some scales?*
> *a) to let the nurse measure her weight*
> *b) to let the nurse measure her body height*
> *c) to measure her weight*
> *d) to measure her body height*

The correct response of #1 was chosen 36% of the time by students who averaged 50.03. Distractor #3 was chosen by 58% of the examinees whose average ability on the test was 51.76. The absolute difference between the levels is only 1.03, so perhaps these students can be judged at a roughly similar level. Distractor #4 was chosen by students of level 55.20, but as the number of students was only three, the possibility that these three students simply slipped up on that item seems likely. Option #2 was selected by 58% of examinees, or 22% more than those who answered correctly. Their average ability was 1.73 points higher. Again, more high ability level examinees answered wrongly. There is very little difference in the wording of both options, the question targets a vocabulary item or phrase. One solution springs to mind. In Japan, some scales have the dual purpose of weighing the body and measuring the height. It is possible that cultural knowledge interfered with examinees ability to separate the meanings in options A and B. These are nursing students under discussion, and even though the non-specialist view of *scales* may be similar in Japanese and English, there remains the possibility that the specialist understanding is different. This can be readily checked by questioning a native Japanese speaker about the semantic space for *scales*. If a discrepancy does exist, future editions of the textbook may need to incorporate it as a teaching point.

Looking at Item 37 again:

> *Q37: It is important to be __________ to new patients.*
> *a) helpful   b) helping   c) helped   d) helper*

On this item, higher ability examinees answered correctly, but the degree of discrimination between those and the others is very narrow, measured at 0.87. The existence of difficult questions in a test does not detract from its validity, but such a fine line is perhaps troubling. The same pattern can be observed for Item 44:

> *Q44: Aerobics is a good way of keeping ___________.*
> *a) exercise   b) fit   c) health   d) lifestyle*

The term *fit* is explained in the textbook, and the adjective-noun distinction *healthy- health* is practiced in the workbook. Options #1 and #4 were dismissed by test takers. These options need to be reworked to allow for a better spread of answer-distractor options. Few examinees selecting an option indicates that that option is not working usefully towards any target the question may have. More usefully, the usages of *healthy* and *fit* may become a teaching point in an updated revision of the textbook.

The absolute measured difference between the 73% of examinees who selected the wrong option and those 14% who answered correctly was 0.21 scaled points. This lack of clear discrimination between levels brings the quality of the test into question. The differentiation between distractors needs to be clearly demarcated, especially that between the correct responses and the others. In this test, most of the items are separated by only a few percentage points.

Conclusions

Both CTT and Rasch indicated some weak items in the test. In the Rasch analysis, Item 27 produced a negative correlation in ID and PT measure values. CTT's IF and ID values identified a number of items that produced questionable figures. None of the IF figures, though, were sufficient to uncategorically eliminate any item. IF provided a clue as to where the problem items were. One by one, an analysis of each item was necessary. ID suggested that there were 12 weakly discriminating items. Winstep's point-measure correlations pointed to over 30 items.

I can prepare the worksheets for IF, ID and split half for a data set of 140+ examinees on 50 questions in about an hour if my template files are available. From scratch, the process would take upwards of two hours. The same amount of data can be used to set up a Winsteps analysis in a few minutes. From then, each analysis requires only a two mouse clicks. As a classroom teacher, the amount of time saved by using Winsteps is considerable. As a materials developer, the readily digestible information is invaluable. However, CTT is conceptually straightforward, while Rasch is not.

In summary, CTT's IF and ID are a good place to begin the analysis of the test. They can indicate potential problems. The key word here is "potential". The analysis needs to go back to the raw data in Excel and hunt for more detailed information. Oftentimes, the trail goes cold as, for example, to discover the exact ID relationships that go beyond the top and the bottom 25% are simply not there. At other times, the search leads back to the original test paper for a study of the actual language in the paper. This is not a bad action, of course, and in all analyses need to end up with the test paper in hand. However, the better the quality of the numerical data, the less the analyst needs to be concerned with items that are not problematic, and the more they can focus on the real issues in the test. In this paper, I have only scratched the surface of what Rasch can do. Its true power lies far outside my current reach. My lack of experience will be clear to specialists reading this; they will have constantly scratched their heads wondering "why didn't he write about this or that?" However, I hope that they may reflect on the gap between their expert position and my own and come forward to help make Rasch more accessible to many who are presently unaware of its might. Rasch provides highly detailed and compelling tools for the analyst. The learning curve, though, is steep.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New Ed.).* New York: McGraw-Hill.

Choppin, B. H. (1983). The Rasch model for item analysis. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

Holster, T. A., & Lake, J. W. (2014). How high can they jump: An introduction to Rasch measurement. *文藝と思想 (Bungei to Shisou: The Bulletin of Fukuoka Women's University International College of Arts and Sciences), 78*, 19-45.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Linacre, J. M. (2014). Winsteps (Version 3.81.0). Retrieved from http://www.winsteps.com

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Sick, J. (2008a). Rasch measurement in language education: Part 1. *Shiken: JALT Testing and Evaluation SIG Newletter, 12*(1), 1-6.

Sick, J. (2008b). Rasch measurement in language education: Part 2. *Shiken: JALT Testing and Evaluation SIG Newletter, 12*(2), 26-31.

Sick, J. (2009). Rasch measurement in language education: Part 3. *Shiken: JALT Testing and Evaluation SIG Newletter, 13*(1), 4-10.

Sick, J. (2010). Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing and Evaluation SIG Newletter, 14*(2), 23-29.

Smiley, J., & Masui, M. (2013). *Nursing Care*. Brighton: Perceptia Press.

From raw scores to Rasch in the classroom

Trevor A. Holster[1] and J. W. Lake[2]
trevholster@gmail.com
1. Fukuoka University
2. Fukuoka Jogakuin University

Abstract

Smiley's experience reported in this issue of *Shiken* is probably quite typical of moving from traditional analysis to Rasch analysis. Traditional analysis, exemplified by Brown's (2005) *Testing in Language Programs*, provides statistics such as item facility values (IF) and item discrimination (ID) which will identify most of the same problematic items as Rasch analysis, and it's unlikely that classroom grades would change to any substantive degree between the two for a thoughtfully developed test. Rasch analysis provides benefits beyond analogues of traditional item analysis, however, and this paper argues that two important practical benefits are the variable map, or Wright map, which provides a quick visual summary comparing students with instructional features, and data-model fit statistics which provide for diagnosis and identification of students requiring remedial instruction. This study illustrates the potential of these for curriculum planning and classroom diagnosis through analysis of the vocabulary section of an academic English placement test.

Keywords: Diagnostic assessment, Rasch, item analysis, vocabulary testing

As Smiley reports in this issue of *Shiken*, traditional item analysis includes item facility values (IF), which rank item difficulty by the proportion of correct responses, and item discrimination (ID), which shows whether high ability persons scored higher overall on an item than low ability persons, a simple assumption being that higher ID is generally better. Rasch software reports several statistics regarding item performance, including point-measure correlation and infit and outfit statistics. As Linacre (2012) explains, the point-measure correlation is closely related to the point-biserial correlation that can be used for the same purpose as ID (Brown, 2005, p. 70), so the closest analogue of ID is the point-measure correlation. Rasch fit statistics are based on a different conception of discrimination, however, and this is fundamental to understanding the differences between the Rasch model and traditional analysis. In traditional ID analysis, we assume that higher scoring students answer correctly more often on all items than lower scoring students, so ID allows us to identify items that behaved unexpectedly. We need some difficult items, i.e. with low IF values, to target high ability students, and these will have high IDs. We also need some easy items, i.e. with high IF values, to target low ability students, and these will have much lower IDs or correlations because many low ability students will answer them correctly. Although negative ID values indicate problematic items, a good test will have items with a range of IF and ID values, so higher ID alone does not automatically indicate a better item. The Rasch model shares the expectation that high ability students will succeed more than low ability students on all items and that point-measure correlations will vary for effective items but should always be positive, but Rasch data-model fit is calculated by comparing the observed discrimination of items, which are never equal, with a theoretical ideal in which all items have equal discrimination (see Sick, 2010, for discussion of Rasch model assumptions). However, Rasch discrimination is very different from the traditional ID value, so traditional analysis has no direct analogue to Rasch fit statistics.

The left hand panel of Figure 1 illustrates this key feature of the Rasch model, showing item characteristic curves (ICCs) for three items of different difficulty. The vertical axis shows the probability of success of a person on an item, ranging from a lower limit of 0.00 to an upper limit of 1.00. The horizontal axis shows person ability in log-odds units, or "logits" (Bond & Fox, 2007). When item difficulty and person ability are perfectly matched, the person has a 50% chance of success, giving odds of 50/50, or 1/1. The natural logarithm of 1/1 equals 0, so an expectation of success of 50% means a difference between item difficulty and person ability of 0.00 logits. In Rasch analysis, there is no absolute zero point indicating zero ability, so 0.00 logits is just an arbitrary point that, by convention, indicates the mean difficulty of the sample of items. In Figure 1, therefore, we would expect 50% of students with ability of 0.00 logits to succeed on an item of average difficulty and 50% to fail. If the same group of students took an easier item, with difficulty of -1.00 logit, we would expect about 73% to pass and about 27% to fail, i.e. odds of about 73/27, because person ability is about 1 logit higher than the item difficulty and the natural logarithm of 73/27 is roughly 1. For a more difficult item of 1 logit difficulty, the probability of success falls to about 27% because odds of 27/73 corresponds to a logit difference of about -1.

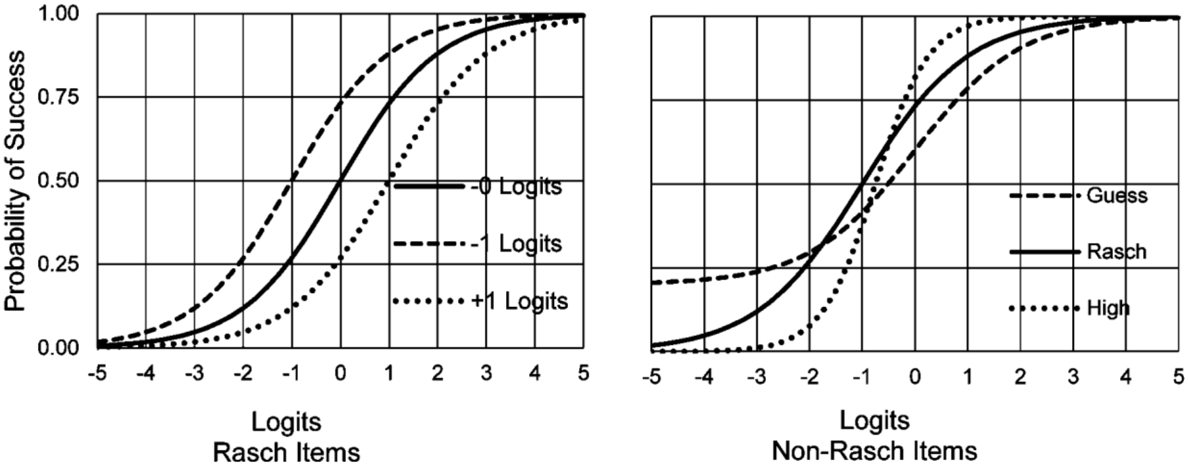

Figure 1. Rasch and non-Rasch item characteristic curves for items of different difficulty levels. The vertical axis shows the probability for three items of different difficulty. The Rasch model assumes parallel curves, but non-Rasch models allow non-parallel curves.

So far this is consistent with the commonsensical expectation that success on items will correlate with person ability, but what is conceptually important about the Rasch ICCs in Figure 1 is that they are parallel, i.e. that the slope of each curve is the same at each point on the vertical axis. The difference in difficulty between the successive items is 1.00 logits at every probability level. In other words, the relative difficulty of the items is theorized to be invariant regardless of the ability of the persons taking the test. Similarly, the relative ability of the persons is theorized to be invariant regardless of the set of items used in the test. The Rasch model thus assumes a stable hierarchy of person ability that does not vary for different samples of items, and a stable hierarchy of item difficulty that does not vary for different samples of persons. This theoretical ideal is only possible if ICCs are parallel (Engelhard, 2013), and, as item discrimination in the Rasch model is simply the slope of the ICC at the 50% expectation of success level, the ideal of invariant measurement is only possible if all items have identical discrimination.

However, real data sets never perfectly fit the idealized Rasch model, in fact, they often misfit quite dramatically. The right-hand panel of Figure 1 shows response patterns that illustrate items that misfit Rasch assumptions. One item, *Rasch* follows a Rasch ICC, but another item, *High*, has a much steeper slope, i.e. higher discrimination. The problem this causes is that low ability students have a higher probability of success on *Rasch* than on *High*, i.e. *Rasch* is easier than *High*, but for high-ability persons, the hierarchy is reversed and *Rasch* is more difficult. This is another way of saying that students seem to have followed different trajectories of acquisition for these two items. In the case of a classroom test where test items are based on course content, if a large number of items misfitted in this way, we might want to investigate to see whether our curriculum is mixing different types of knowledge and skills that should be assessed separately. Another source of misfit is shown by the item *Guess*. In this case, the expectation of success does not approach the lower limit of zero assumed by the Rasch model, so even very low ability persons still have about a 20% chance of success. This is the type of pattern we might see in a situation where guessing is possible, such as a multiple-choice test with five answer choices or in a constructed response question that gives clues to the answer. An important point about these three items is that the major problem for the Rasch model is mixing items with ICCs that diverge too much from parallel trajectories, so items that function well in one test may misfit if used in a test that measures a different type of knowledge.

Rasch analysis provides a simple diagnostic tool to identify items or persons that violate the model's assumptions in the form of mean-square fit statistics. Fit statistics are generated from the patterns of the differences between observed responses and statistically expected responses, known as score residuals. In a dichotomously scored test, observed responses can only have values of 0 or 1, while expected responses, or probability of success on items, can take any value between the asymptotes of 0.00 and 1.00, so observed values and expected values can never be exactly equal. When person ability and item difficulty are perfectly matched, the probability of success equals .50, so the residual is 0.50 for a correct answer and -0.50 for an incorrect answer. Small residuals will occur when high ability persons succeed on easy items or low ability persons fail on difficult items, while large residuals will occur when low ability persons succeed on difficult items or high ability persons fail on easy items. Across the entire data set, these values are expected to follow a chi-square distribution, and the mean-square fit statistics provide a confirmatory analysis to see whether the observed data fit the modelled distribution.

The mean-square statistic has an expected value of 1.00, indicating patterns of responses that perfectly match the Rasch model, with a lower limit of zero and no upper limit. Mean-square values below 1.00 indicate responses that are more predictable than expected, called overfit, while mean-square values greater than 1.00 indicate less predictable responses, called misfit (or underfit). For the three items in the right hand panel of Figure 1, *Rasch* would show perfect data-model fit, but *High* would overfit the model and *Guess* would misfit the model. In the real world, some items and persons will inevitably be more consistent than average and some will be less consistent, so aiming for perfect data-model fit is not the objective. Rather, we need to investigate whether the misfit is severe enough to threaten the interpretations we wish to make of the test scores and whether there are systematic patterns of misfit that indicate sampling problems with either items or persons.

While much of the published research on language testing is from the perspective of large-scale standardized proficiency tests, where practicality and reliability are paramount concerns, a materials writer or textbook planner who wishes to integrate assessments into a course of study may be more concerned with criterion referencing student ability against instructional content or in diagnosing students or instructional items that follow unusual developmental trajectories. Rasch analysis provides useful tools for this, so the purpose of this paper is to demonstrate its benefits within instructional programs. This study was conducted as part of a curriculum development project for reading classes in a newly established Academic English Program (AEP) at a Japanese public woman’s university. In 2011, the first year of the program, detailed goals and objectives were not available and different teachers used different reading textbooks and instructional approaches. Students' TOEFL score trajectories diverged from the assumptions of the university and prefecture, so textbook selection was reviewed and a placement test developed for 2012, with intended secondary uses as a diagnostic and achievement test. The 2012 test form had three sections of 50 items each and this was revised in 2013 to five sections of 40 items each, including content derived from the assigned textbook series, *Reading Explorer* (Douglas & McIntyre, 2009).

Each of the five different levels in the *Reading Explorer* series comprised 24 reading passages followed by five comprehension check questions intended to prepare students for tests such as the TOEFL. Each reading passage targeted 10 academic words for explicit instruction, but it was apparent that many students needed to study non-academic words in the textbook as well, so supplementary vocabulary instruction was required and a vocabulary test was needed to determine appropriate vocabulary for students of different proficiency. Given the TOEFL orientation of the program, Davies and Gardner's *A Frequency Dictionary of Contemporary American English* (2010) was adopted as the basis of the vocabulary section of the placement test, on the assumption that higher frequency words are generally more important to learn and more likely to be integrated into long-term knowledge because they will tend to be encountered more frequently in authentic use. Thus, Section 1 of both test forms aimed to measure vocabulary knowledge at different levels of word frequency. Each frequency band of 1000 words from Davies and Gardner (2010) was tested by 10 items, with the expectation that the average difficulty of items would increase as word frequency decreased, allowing the lexical burden of reading passages to be estimated for students at different TOEFL levels, providing evidence to guide textbook selection for classes of different levels.

> RQ1. Did the difficulty of items in the vocabulary section follow the hypothesized hierarchy based on word frequency?
>
> RQ2. Did students demonstrate good data-model fit, indicating that students from different high-schools followed similar trajectories of vocabulary learning?

Method

Participants

All participants were female Japanese university students enrolled in an academic English program at a public Japanese women's university. Placement tests were administered in April 2012 and April 2013 to assign students to both academic English classes and first-year seminar classes conducted in Japanese. The 2012 cohort had 249 students and the 2013 cohort had 243 students, for a total of 492 students.

Instrument

The vocabulary section of the test comprised 50 items in 2012 but was reduced to 40 items in 2013. Although the VST (Beglar, 2010; Nation & Beglar, 2007) was considered as a source of vocabulary items, the frequency lists provided by Davies and Gardner (2010) were considered to more relevant to the AEP’s academic focus so a new test was developed using a synonym matching format instead of the definitions used in the VST. As the students were enrolled in an academic English program, knowledge of very high frequency vocabulary was assumed, so each item stem used a word taken from the first 500 words listed by Davies and Gardner (2010), with the correct answer, the key, being synonymous with this. The distractors were of similar frequency to the key and of the same part of speech, so the difficulty of items was hypothesized to result from the frequency of the key and distractors. A sample item is shown below:

> With
> A) Ago B) Least C) Enough D) Already E) Together

For the 2012 test, 10 items were sampled from each of the 1000 word frequency bands in Davies and Gardner's 5000 word list, giving 50 items in total. Analysis of the 2012 results showed that the items from the first and second 1000 frequency bands (1K and 2K) were too easy for most students, so these were replaced with 10 academic items derived from the reading textbook series for the 2013 test, leaving 40 items. This analysis therefore includes 60 items, with 1K and 2K items used only in 2012, academic items used only in 2013, and 3K, 4K, and 5K items linking the two subsets of data.

Procedure

Tests were administered on the first day of semester, supervised by AEP teachers. Administrative constraints dictated a two-hour time limit for the placement test, raising concern over speededness. Observation during test administration showed that most students finished all sections within the allotted time, and that speededness did not affect the vocabulary section, which was the first section of both test forms. Therefore, missing responses were coded as incorrect. All analyses were conducted using Winsteps (Linacre, 2010). Following each test administration test forms were scanned and processed using Remark Office OMR version 8.4 (Gravic, 2012), response data exported to Microsoft Excel 2010 (Microsoft, 2010), and then imported into Winsteps as a plain text file.

Results

Figure 2 shows the variable map, or Wright map, with mean item difficulty set as 400 and 1 logit scaled to 50 points, to give an approximation to the TOEFL scale. The vertical scale thus allows visual comparison of person ability and item difficulty because both are measured in the same units. Persons are shown in the left column, with item distribution shown in the second column, and items displayed by frequency band in the rightmost six columns. Items are labeled by frequency level, with targeted academic vocabulary labeled as "AW". Most persons fall within the TOEFL 400 to 500 range, consistent with students being unable to read unsimplified texts upon entry to the AEP. Students with TOEFL levels of 400 would have a 50% expectation of answering an average item, while students at the 500 level would have an 88% expectation of success on an average item. A general trend can be seen for high-frequency items to be easy and academic items to be difficult, but the pattern is not strongly deterministic, with one very easy 5K item and two very difficult 2K items. This is supported by Table 1, showing mean item difficulty by frequency band. This must be interpreted very cautiously because 10 items per frequency band is insufficient for definitive results, but Table 1 shows the expected trend of mean item difficulty increasing as vocabulary decreases. Although adjacent frequency bands aren't clearly separated, with some 3K words easier than most 1K and 2K words, the evidence supports the view that Davies and Gardner (2010) provide a useful classroom guide for prioritizing vocabulary items.

However, a curriculum aims to match students with language features of appropriate difficulty, and this is where the benefits of Rasch measurement become apparent. Although raw scores can rank-order person ability and item difficulty, they do not place person ability and item difficulty on a shared measurement scale. Also, rank-ordering using raw scores requires that all persons take the same set of items and that all items are taken by the same set of persons unless equating procedures are used, greatly complicating matters when different test forms are used, as in this study. Rasch analysis provides comparison of both persons and items on the same measurement scale even when different test forms are used, as long as the test forms have a subset of 10 or more common items that can be employed to statistically link the forms, so the Wright map shown in Figure 2 allows curriculum planners and classroom teachers to quickly see the relative ability of each student compared to instructional items. We can see that very few persons were below 400, but 1K and 2K words mostly fell below this, while academic words were mostly above the average person ability of about 470. Therefore, it seems reasonable to focus vocabulary instruction on 3K and 4K words for most students, while reviewing and consolidating 1K and 2K words with the lowest group and introducing academic words with the upper levels. In this way, Rasch analysis allows curriculum planners and materials writers to visually compare the levels of students and items to check that instruction is appropriately targeted.

However, as Figure 1 showed, the hierarchy of item difficulty of the Wright map assumes adequate data-model fit. The "pathway" maps produced by Winsteps provide a simple visual tool to investigate this. Figure 3 shows the pathway maps for items, shown in the left-hand panels, and persons, shown in the right-hand panels. The vertical scale shows item difficulty and person ability. Each circle represents one item or one person, with the size of the circle representing the measurement error. If two circles overlap on the vertical scale, we do not have 95% confidence that they are different in difficulty or ability. The horizontal scale shows mean-square fit statistics, which can range from zero to infinity. A value of 1.00 indicates that the randomness in the data matches the expectations of the Rasch model, while values below this indicate unexpectedly predictable data and values higher than 1.00 indicate noisy data. Linacre suggests a rule-of-thumb that mean-square statistics between 0.5 and 1.5 are productive of measurement. However, two sets of mean-square statistics are produced, information weighted infit statistics, shown in the upper panels, and unweighted outfit statistics, shown in the lower panels. The information weighting of the infit statistic emphasizes responses where person ability and item difficulty are well matched because this is where information is maximized, so this statistic is the crucial indicator of whether the instrument supports measurement. The outfit statistic, generated from unweighted responses, shows the effect of outlying responses, such as when low ability persons succeed on difficult items or high-ability persons fail on easy items. Comparison of the patterns of infit and outfit thus gives important diagnostic information about where unexpected responses are occurring.

```
Scale|                   Persons | All Items || Items by Frequency Band            |Scale
     |                           +           ||  1K |  2K |  3K |  4K |  5K |  AW  |
 600 |                           +T          ||     |     |     |     |     |      | 600
 590 |                           +           ||     |     |     |     |     |      | 590
 580 |                        .  +           ||     |     |     |     |     |      | 580
 570 |                        .  +  *        ||     |     |     |     |  5  |      | 570
 560 |                      .** T+  *        ||     |     |     |     |     |  A   | 560
 550 |                       .*  +  *        ||     |  2  |     |     |     |      | 550
 540 |                     .***  +  **       ||     |     |     |  4  |     |  A   | 540
 530 |                  .******  +  *        ||     |     |     |     |     |  A   | 530
 520 |               .*********  +  *        ||     |     |  3  |     |     |      | 520
 510 |                .******** S+  **       ||     |     |     |     |     +M AA  | 510
 500 |             .***********  +S ****     ||     |  2  |     |     |  55 |  A   | 500
 490 |     .*******************  +  *        ||     |     |     |     |     |  A   | 490
 480 |         .***************  +  ***      ||     |     |     |  4  |  5  |  A   | 480
 470 |          .************** M+  *        ||     |     |     |     |     |  A   | 470
 460 | .***********************  +  ***      ||     |     |     |  4  +M 5  |  A   | 460
 450 |                 ********* +  **       ||  1  |     |     |     |  5  |      | 450
 440 |               .*********  +  ****     ||     |     |  3  |  44 |  5  |      | 440
 430 |                 ******** S+  *        ||     |     |     +M 4  |     |      | 430
 420 |                     .***  +  *        ||     |     |     |  4  |     |      | 420
 410 |                      **** +  ***      ||     |     |  333|     |     |      | 410
 400 |                       .*  +M          ||     |     +M    |     |     |      | 400
 390 |                        ** +  *        ||     |  2  |     |     |     |      | 390
 380 |                       .* T+  ***      ||  1  |     |  3  |     |  5  |      | 380
 370 |                        ** +  **       ||     |     |     |  4  |  5  |      | 370
 360 |                         * +  *        ||     |  2  |     |     |     |      | 360
 350 |                        .  +  *        ||     |     |     |  4  |     |      | 350
 340 |                        .  +  **       ||     +M 2  |  3  |     |     |      | 340
 330 |                       .*  +  **       ||  1  |  2  |     |     |     |      | 330
 320 |                           +  *        ||     |  2  |     |     |     |      | 320
 310 |                        .  +  ***      |+M 111|     |     |     |     |      | 310
 300 |                        .  +S          ||     |     |     |     |     |      | 300
 290 |                           +  ***      ||  1  |  2  |     |     |  5  |      | 290
 280 |                           +           ||     |     |     |     |     |      | 280
 270 |                           +  ****     ||  1  |     |  33 |  4  |     |      | 270
 260 |                           +  *        ||  1  |     |     |     |     |      | 260
 250 |                           +           ||     |     |     |     |     |      | 250
 240 |                           +           ||     |     |     |     |     |      | 240
 230 |                           +  **       ||     |  2  |  3  |     |     |      | 230
 220 |                           +  *        ||  1  |     |     |     |     |      | 220
 210 |                           +           ||     |     |     |     |     |      | 210
 200 |                           +T*         ||     |  2  |     |     |     |      | 200
     |                           +           ||  1K |  2K |  3K |  4K |  5K |  AW  |
Scale|                   Persons | All Items || Items by Frequency Band            |Scale
Note: Each "*" in the person column is 3 persons, each "." is 1 to 2
```

Figure 2. Person-item map showing person ability and item difficulty scaled to approximate TOEFL level. Items are identified by frequency band, with academic items identified as "A". "M" indicates the mean difficulty for all items, and the median for each frequency band.

Table 1
Item Difficulty by Frequency Band

		Item Difficulty			
Item Level	Count	Mean	Median	S.D.	S.E. Mean
1K	10	311.74	307.31	61.36	20.45
2K	10	351.23	337.51	102.94	34.31
3K	10	368.69	396.05	83.45	27.82
4K	10	418.75	432.46	71.51	23.84
5K	10	442.98	458.18	75.22	25.07
AW	10	506.62	505.19	30.13	10.04
All	60	400.00	412.22	98.08	12.77

Note: Subtotal reliability =.85
Scale: Mean item difficulty = 400, 1 logit = 50

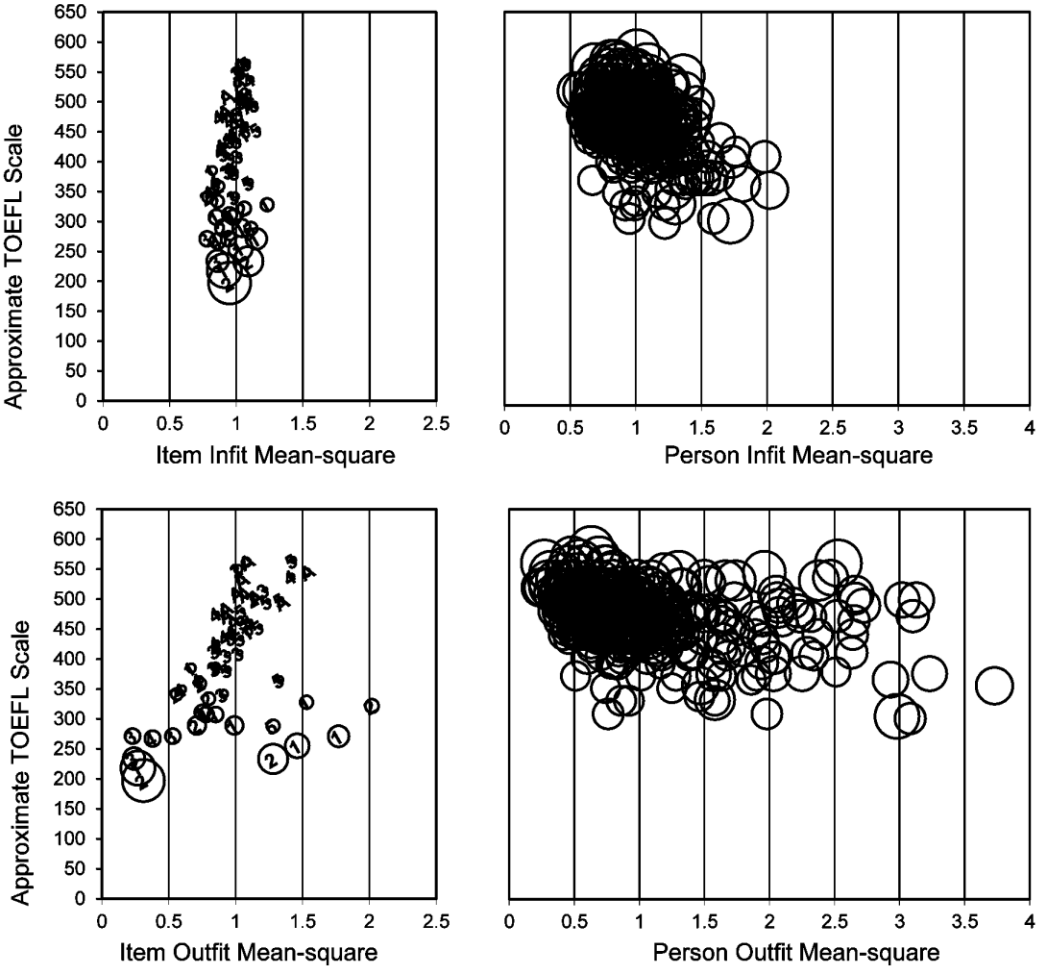

Figure 3. Pathway maps showing data-model fit. The vertical axis shows an approximate TOEFL scale. Each bubble shows a single item or person, with the size of the bubble indicating an approximate 95% confidence band of difficulty or ability. The horizontal axis shows mean-square fit statistics, with 1.5 being a rule-of-thumb threshold for concern.

In the case of the items, the infit statistics are extremely good, but the outfit statistics show several misfitting items that did not follow the parallel acquisition trajectories assumed by the Rasch model. The person statistics show a more worrisome pattern, with the infit statistics showing a number of misfitting persons and the outfit statistics showing many. Thus, many students are not displaying parallel trajectories of vocabulary acquisition, suggesting idiosyncratic exposure to English vocabulary at high-school or from studying for university entrance exams. Although the item statistics indicate a relatively stable hierarchy of item difficulty, the evidence points to many students having idiosyncratic vocabulary knowledge. This suggests the need for remedial instruction for higher ability students who incorrectly answered easy items, and thus might struggle with high-frequency vocabulary despite having considerable knowledge of academic vocabulary. Winsteps provides an accessible solution to this in the form of Kidmaps.

Figure 4 shows the Kidmap for one student. The central vertical scale shows item ability, with the student's ability estimated as 499 plus or minus 22 and the horizontal bars showing the 95% confidence band. The left-hand side of the map identifies the items that were answered correctly, while the right-hand side identifies the items answered incorrectly, so "35.1" indicates Item 35 was given a score of 1, while "33.0" indicates that Item 33 was given a score of 0. The items in the lower right quadrant show unexpected failures so this student should revise Items 33, 9, 27, 50, 45, and 43. What is notable is that this student is above average in ability but has followed an acquisition trajectory that diverges from the overall group, so the remediation is targeting easy items that were incorrect. In contrast to remediation aimed at helping low ability students close the gap to average students, this remediation targets higher ability students with idiosyncratic knowledge in order to bring them in line with the sequencing assumed by the curriculum planner. In contrast, the upper right quadrant shows expected failures, so this provides a sequence for non-remedial instruction of items above the student's current level.

```
Name: Student
Measure: 498.54  S.E. 21.99  Score: 41
Test: Test of Vocabulary Synonymity

Correct Answers       -Harder-      Incorrect Answers
-----------------------------------------------------
|                        600                        |
|                         |                         |
|                         |                         |
|                         | 47.0                    |
|                         | 14.0                    |
| 35.1                   540                        |
|                         |                         |
|---------------------------30.0--------------------|
| 18.1  44.1             XXX                        |
| 41.1                    |                         |
|-34.1-------------------480-43.0-------------------|
| 38.1                    | 45.0                    |
| 46.1                    |                         |
| 3.1  39.1               |                         |
| 21.1  36.1  37.1        | 50.0                    |
| 28.1  31.1             420                        |
| 29.1                    | 27.0                    |
|                         |                         |
| 7.1  12.1  24.1  42.1   |                         |
|                         |                         |
| 17.1  32.1  49.1       360                        |
| 40.1                    |                         |
| 13.1  20.1  25.1        |                         |
| 19.1                    | 9.0                     |
| 2.1  4.1  5.1           |                         |
|                        300                        |
| 6.1  16.1  48.1         |                         |
| 10.1  23.1  26.1        |                         |
|                         | 33.0                    |
| 8.1                     |                         |
|                        240                        |
| 15.1  22.1              |                         |
| 1.1                     |                         |
|                         |                         |
| 11.1                    |                         |
|                        180                        |
-----------------------------------------------------
Correct Answers       -Easier-      Incorrect Answers
```

Figure 4. Diagnostic Kidmap for a single student showing correct and incorrect responses by item difficulty. Logit measures are shown on the vertical scale. The upper left quadrant shows items with unexpected success, indicating items requiring investigation, and the lower right quadrant shows items with unexpected failure, indicating items requiring remedial instruction.

Discussion and Conclusions

This study aimed to demonstrate how Rasch analysis can be of practical value to curriculum planners, materials writers, and classroom teachers using data from the vocabulary section of an academic English placement test. Although traditional analysis of raw scores can rank-order item difficulty and person ability, and techniques are available to criterion reference person ability to language features, Rasch analysis provides an extremely practical solution to these, while Rasch data-model fit provides a simple conceptual framework for diagnostic assessment. The key theoretical assumption of the Rasch model is that all items and persons follow parallel developmental trajectories and mean-square fit statistics provide an indication of the magnitude of deviations from this idealization. In this study, items showed acceptable data-model fit, supporting the existence of a stable hierarchy of item difficulty. The Wright map is emblematic of Rasch analysis and visually confirmed the hypothesized trend for item difficulty to increase as vocabulary frequency decreased. This provided a practical guide to inform teachers about vocabulary that is likely to cause difficulty for students of different levels. However, many students misfitted the Rasch model, suggesting idiosyncratic trajectories of vocabulary acquisition in high-school English classes and supporting the need for remedial instruction for high-level students with misfitting response patterns. The Kidmap produced by Winsteps provided an individualized diagnostic report to identify test items requiring remediation. These findings illustrate that Rasch analysis has benefits for language programs beyond the identification of misbehaving items, providing insights into the behavior of individual students that are conceptually simple enough for non-specialists to interpret.

Acknowledgments

This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research Grant #25370643. We are also very grateful to Jim Sick for his valuable comments and feedback.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118. doi: 10.1177/0265532209340194

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New Ed.).* New York: McGraw-Hill.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. New York: Routledge.

Douglas, N., & McIntyre, P. (2009). *Reading explorer*. Boston: Heinle.

Engelhard, G. (2013). *Invariant measurement*. New York: Routledge.

Gravic. (2012). Remark Office OMR (Version 8.4).

Linacre, J. M. (2010). Winsteps (Version 3.70.02). Retrieved from http://www.winsteps.com

Linacre, J. M. (2012). A User's Guide to Winsteps, Minstep Rasch-Model Computer Programs Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Microsoft. (2010). Excel (Version 2010).

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Sick, J. (2010). Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing and Evaluation SIG Newletter, 14*(2), 23-29.

Questions and answers about language testing statistics:

Testing intercultural pragmatics ability

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

What sorts of tests have been developed and used for testing intercultural pragmatics ability? What do we know about such testing? And, how have those tests been analyzed statistically?

Answer:

The literature on developing intercultural pragmatics tests has (a) found that different testing formats vary in their effectiveness for testing pragmatics, (b) discovered that certain variables are particularly important in testing pragmatics tests, and (c) relied on increasingly sophisticated statistical analyses in studying pragmatics testing over the years. I will address each of these three issues in turn.

Different Testing Formats Vary in Their Effectiveness for Testing Pragmatics

Starting with Hudson, Detmer, and Brown (1992, 1995), six testing methods have been prominent to varying degrees in the literature to date (as shown in Table 1):

- *Multiple-choice Discourse Completion Task* (MDCT) – requires examinees to read a situation description and choose what they would say next.
- *Oral Discourse Completion Task* (ODCT) – expects examinees to listen to an orally described situation and record what they would say next.
- *Discourse Role-Play Task* (DRPT) – directs examinees to read a situation description and then play a particular role with an examiner in the situation.
- *Discourse Self-Assessment Task* (DSAT) – asks examinees to read a written description of a situation and then rate their own pragmatic ability to respond correctly in the situation.
- *Role-Play Self-Assessment* (RPSA) – instructs examinees to rate their own performance in the recording of the role play in the DRPT.

Hudson et al. (1992, 1995) created initial prototype tests and validated them for EFL students at a US university. They noted that the MDCT did not work particularly well for them. Yamashita (1996) then created Japanese versions of those same tests and verified that all but MDCT worked reasonably well for Japanese as a second language (SL). Enochs and Yoshitake (1996) and Yoshitake (1997) verified that the six assessments worked well for Japanese university EFL students. Ahn (2005) created Korean versions for all but the MDCT and verified that they worked reasonably well for Korean as a FL. Liu (2007) reported on developing a MDCT that worked, which he accomplished by using students to generate the speech acts and situations that were used.

Hudson et al. (1992, 1995) and a majority of the other researchers have used paper-and-pencil testing formats. However, other formats have also been used. Tada (2005) was the first to create computer-delivered tests with video prompts. Roever (2005, 2006, 2007, 2008) was the first to develop and use web-based testing followed by Itomitsu (2009). Rylander, Clark, and Derrah (2013) focused on the importance of video formats. And, Timpe (2013) was the first to use *Skype* role-play tasks.

Certain Variables Are Particularly Important in Testing Pragmatics

In creating their first prototype tests, Hudson et al. (1992, 1995) identified a number of variables that have proven important across many of the subsequent studies, but to varying degrees. These variables are labeled across the top of Table 1. The first was the six ***testing methods*** discussed in the previous section. The second variable was ***speech acts***, which initially included three key ones: (a) *requesting* (i.e., asking another person to do something or for something), (b) *refusing* (i.e., rejecting another person's request), and (c) *apologizing* (i.e., acknowledging fault and showing regret for doing or saying something). The third variable was ***contextual conditions***, which initially included three key conditions: (a) *imposition* (i.e., the degree of inconvenience to the listener of the request, refusal, or apology), (b) *power difference* (i.e., the degree and direction of differences in power or position between the speaker and listener), and (c) *social distance* (i.e., the degree of shared social familiarity or solidarity between the speaker and listener).

Other variables were added as research continued. For example, Roever (2005, 2006, 2007, 2008) added the assessment of idiosyncratic and formulaic implicatures, as well as situational routines in addition to speech acts. He also added rejoinders after the response slot in designing his items. Tada (2005) specifically examined perception versus production of pragmatics to his study. Liu (2006, 2007) innovatively used speech acts and situations generated by students. Grabowski (2009, 2013) examined the relationship between grammar and pragmatic knowledge (which he further subdivided into sociolinguistic, sociocultural, psychological knowledges). Itomitsu (2009) also studied grammar and three aspects of pragmatics (appropriate speech acts, routines, and speech styles) and used requests speech acts, but also added offers and suggestions. Roever (2013) focused on implicature, but also considered vocabulary, collocations, idiomatic word meanings, and morphology. Rylander et al. (2013) added a number of speech acts using refusals and apologies, but also compliments, farewells, greetings, introductions, invitations, suggestions, offers, and complaints. Timpe (2013) included new speech acts: in addition to requests, she used offers, and also examined routine phrases, and phrases/idioms. Youn 2013 added speech acts of expressing opinion and giving feedback on email and compared role-plays with monologic speaking and pragmatics tasks. And finally, Youn and Brown (2013) compared heritage and non-heritage KFL students' performances on such tests.

Increasingly Sophisticated Statistical Analyses have Been Used to Study Pragmatics Tests

A quick glance at the second to last column in Table 1 will reveal that all of the studies have used classical testing theory (CTT), which involves traditional descriptive statistics, reliability estimates, correlation coefficients, and in some cases item analyses. However, as time went by, researchers increasingly used three more complex analyses:

- *Rasch analysis* allows researchers to put items and examinees on the same logit scales.
- *FACETS analysis* is a variation of Rasch analysis that allows researchers to put a variety of different facets (e.g., items, raters, rating categories, etc.) on the same logit scale and, among other things, allows simultaneous display of whatever facets are selected so they can be compared to examinee performances (for instance, examinees can be represented on the same scale as raters and rating categories, as in Brown, 2008).

Table 1

Pragmatics Testing Projects (Quantitative) Described in Terms of Testing Methods, Speech Acts, Contextual Conditions and Value Added to the Knowledge of Pragmatics Assessment[1]

	Testing Method[1]						Speech Acts				Conditions[2]					
Testing Project; L2 Being Learned and where	WDCT	MDCT	ODCT	DRPT	DSAT	RPSA	Requesting	Refusing	Apologizing	Other	Imposition	Power	Social Dis.	Test Type[3]	Statistical Analyses[4]	Value Added to Knowledge of Pragmatics Assessment
Hudson et al. 1992, 1995; ESL in US	X	X	X	X	X	X	X	X	X		2	2	2	P&P	CTT	Created the initial tests and validated all but the MDCT for EFL students at a US university.
Enochs & Yoshitake, 1996; Yoshitake 1997; Both EFL in Japan	X	X	X	X	X	X	X	X	X		2	2	2	P&P	CTT	Verified that six assessments worked reasonably well for Japanese university EFL students; scores also compared to 3 TOEFL subtests.
Yamashita 1996; JSL in Japan	X	X	X	X	X	X	X	X	X		2	2	2	P&P	CTT	Created Japanese versions and verified that all but MDCT worked reasonably well for Japanese as a SL.
Ahn, 2005; Brown 2008; Brown & Ahn, 2011; All KFL in US	X		X	X	X	X	X	X	X		2	2	2	P&P	CTT, G theory, FACETS	Examined the effects of numbers of raters, functions, item types, and item characteristics on reliability and difficulty/severity in several combinations.
Roever 2005, 2006, 2007; ESL/EFL in US/Germany/Japan		S					X	X	X		2	1	1	WBT	CTT, FACETS, DIF	Assessed idiosyncratic and formulaic implicatures, situational routines, and speech acts; formats similar to MDCT, but speech acts added rejoinders after the response slot.
Tada 2005; EFL in Japan		S	S				X	X	X		2	2	1	CLT, Video	CTT	1st to be computer delivered with video prompts for tests similar to MDCT and OPDCT (specifically examined perception vs. production of pragmatics)
Liu 2006; EFL in PRC	S	S			S						2	2	2	P&P	CTT, Rasch	Speech acts and situations were generated by students.
Liu 2007; EFL in PRC	S										2	2	2	P&P	CTT, Rasch	Focused on developing a MDCT that worked; Speech acts and situations were generated by students.
Roever 2008; ESL/EFL in US/Germany/Japan		S					X	X	X		2	1	1	WBT	CTT, FACETS	Speech acts section only; rejoinders after the response slots; examined effects of raters and items.
Youn 2008; KFL in US	X		X	X			X	X	X		2	2	2	P&P	CTT, FACETS	Examined the effects of test types and speech acts on raters assessments.
Grabowski 2009, 2013; ESL in US				S								1	2	P&P	CTT, G theory, FACETS	Speaking tests similar to DRPT; rated and examined the relationship between grammar and pragmatic knowledge (further subdivided into sociolinguistic, sociocultural, psychological knowledges).
Itomitsu, 2009; JFL in US		S					X			X				WBT	CTT	Grammar and three aspects of pragmatics (appropriate speech acts, routines, and speech styles); three not distinguishable; only total scores validated; speech acts included requests, offers, suggestions.
Roever, 2013; NS & ESL in Australia		S												P&P	CTT, FACETS	Focuses on implicature (along with subtests on vocabulary, collocations, idiomatic word meanings, & morphology)
Rylander, Clark, & Derrah, 2013; EFL in Japan							X		X	X				P&P, Video	CTT, Rasch	Focuses on importance of video: added speech acts (refusals & apologies , but also compliments, farewells, greetings, introductions, invitations, suggestions, offers, & complaints).
Timpe, 2013; EFL in Germany		S			S		X			X		2	2	WBT	CTT, Rasch	Focused on American English self-assessment, a sociopragmatic comprehension test, and *Skype* role-play tasks. Sociopragmatics test include speech acts (requests and offers), routine phrases, and phrases/idioms
Youn 2013; KFL in US				S			X			X		2		P&P	CTT, FACETS	(a) based on needs analysis, developed open role-play tasks similar to DRPT but more interactive; (b) added speech acts of expressing opinion and giving feedback on email; (c) compared role-play with monologic speaking and pragmatics tasks; &(d) exceptionally thorough reliability & validity study based on Kane's (2006) argument-based approach.
Youn & Brown, 2013; KFL in US	X		X	X			X	X	X		2	2	2	P&P	CTT, FACETS	Focused on comparison of heritage and non-heritage KFL students

[1] X = adapted same test; S = Similar test

[2] Number of levels (1 or 2) of each condition, e.g, Imposition high or low would be 2 levels

[3] P&P = Paper & Pencil test; CLT = Computerized Language Testing; WBT = Web-based Language Testing

[4] CTT = Classical Test Theory; G-theory = Generalizability theory; Rasch = Rasch analysis; FACETS = Multifaceted Rasch analyses; DIF = Differential Item Functioning

[1] Only quantitative research studies are considered here. In addition, whenever multiple publications appeared to be based on the same data, I grouped them as one project.

- *Generalizability theory* (G theory) allows researchers to study and minimize multiple sources of error in two stages: (a) a *Generalizability study*, which is used to estimate variance components for whatever facets the researcher wishes to study and thereby to understand the relative proportions of variance accounted for by the object of measurement (usually variance due to examinees) and other facets that are sources of variance (for example, raters and rating categories) (note that this can be done for either norm-referenced or criterion-referenced tests by using different procedures) and (b) a *Decision study*, which is used to estimate the appropriate generalizability coefficients (analogous to reliability estimates) for different numbers of levels in each facet (e.g., estimates can be provided for 2 raters or 3, 4, 5, etc. while also examining what happens simultaneously if 2 rating categories are used or 3, 4, 5, 6, etc.). For an example of this entire process, see Brown and Ahn (2013).

These analyses and others have been applied in various ways with generally increasing levels of sophistication in the pragmatics testing literature. Hudson et al. (1992, 1995) created the initial tests and validated all but the MDCT for EFL students at a US university using CTT. Enochs and Yoshitake (1996) and Yoshitake (1997) verified that the six assessments worked reasonably well for Japanese university EFL students using CTT. Those scores were also compared to the three sets of TOEFL subtest scores available at that time. Yamashita (1996) created Japanese versions and verified that all but MDCT worked reasonably well using CTT. Ahn (2005), Brown (2008), and Brown and Ahn (2011) used FACETS and G-theory analyses to examine the effects of numbers of raters, functions, item types, and item characteristics on reliability and difficulty/severity in various combinations. Roever (2005, 2006, 2007) used FACETS and differential item functioning analyses. Liu (2006) used Rasch analysis to study the effectiveness of speech acts and situations that had been generated by students. Liu (2007) also used Rasch analysis but focused on developing a MDCT that worked. Roever (2008) applied FACETS analysis to study the effects of raters and items. Youn (2008) used FACETS analysis to examine the effects of test types and speech acts on raters assessments. Grabowski (2009, 2013) used both G theory and FACETS analysis in the process of examining speaking tests similar to DRPT with a focus on the relationship between grammar and pragmatic knowledge. Roever (2013) used FACETS analysis in his study of implicature. Rylander et al. (2013) used Rasch analysis in their study testing many different speech acts while using video formats. Timpe (2013) also used Rasch analysis in her study of American English self-assessment, a sociopragmatic comprehension test, and *Skype* role-play tasks. Youn (2013) relied on Rasch analysis in her elaborate validity study (based on Kane's (2006) argument-based approach) of role-plays with monologic speaking and pragmatics tasks. And finally, Youn and Brown (2013) used FACETS analysis in their comparison of heritage and non-heritage KFL students' performances.

Conclusion

Different testing formats (including the original WDCT, MDCT, ODCT, DRPT, DSAT, RPSA, and a number of variations on those themes) have been shown to vary in their effectiveness for testing pragmatics depending on the context and the variables involved. In the process, a wide range of variables have been studied in the literature to date (especially, testing methods, speech acts, and various conditions). In addition, CTT, Rasch, FACETS, and G theory have been the major forms of analysis in the increasingly sophisticated pragmatics testing literature in a variety of different ways.

In all probability, pragmatics testing will continue to grow in the future. No doubt additional tests will be developed (a) to assess pragmatics in additional languages, (b) to accommodate new additional variables as the subfield of intercultural pragmatics continues to expand, and finally, (c) to adjust to refinements in pragmatics constructs and testing formats. It will be interesting to see what impacts all this activity will have on the teaching and testing of English and other languages around the world—and of course in Japan.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University.

Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301-325). Cambridge: Cambridge University.

Brown, J. D., & Ahn, C. R. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics, 43*, 198-217.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Grabowski, K. C. (2013). Investigating the construct validity of a role-play teste designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 149-171). London: Palgrave Macmillan.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Technical Report #7). Honolulu: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Itomitsu, M. (2009). *Developing a test of pragmatics of Japanese as a foreign language*. Unpublished Ph.D. dissertation, Columbia University.

Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners*. Frankfurt am Main: Lang.

Liu, J. (2007). Developing a pragmatic test for Chinese EFL learners. *Language Testing, 24*, 391-415.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165–189.

Roever, C. (2008). Rater, item, and candidate effects in discourse completion tests: A FACETS approach. In E. Soler & A. Martinez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching, and testing* (pp. 249-266). Bristol, UK: Multilingual Matters.

Roever, C. (2013). Testing implicature under operational conditions. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 43-64). London: Palgrave Macmillan.

Rylander, J., Clark, P., & Derrah, R. (2013). A video-based method of assessing pragmatic awareness. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 65-97). London: Palgrave Macmillan.

Tada, M. (2005). *Assessment of ESL pragmatic production and perception using video prompts*. Unpublished doctoral dissertation, Temple University, Japan.

Timpe, V. (2013). *Assessing intercultural language learning.* Frankfurt am Main: Lang.

Yamashita, S. O. (1996). *Six measures of JSL Pragmatics* (Technical Report #14. Second Language Teaching and Curriculum Center). Honolulu, HI: University of Hawai‘i.

Yoshitake, S. S. (1997). *Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation.* Unpublished doctoral dissertation, Columbia Pacific University, Novata, CA.

Youn, S. J. (2008). *Rater variation in paper vs. web-based KFL pragmatic assessment using FACETS analysis.* Unpublished MA thesis, University of Hawai‘i at Mānoa.

Youn, S. J. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods.* Unpublished PhD dissertation, University of Hawai‘i at Mānoa.

Youn, S. J., & Brown, J. D. (2013). Item difficulty and heritage language learner status in pragmatic tests for Korean as a foreign language. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 98-123). London: Palgrave Macmillan.

Where to Submit Questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA