Examining the reliability of a TOEIC Bridge practice test under 1- and 3-parameter item response models

Jeffrey Stewart
jeffjrstewart@gmail.com
Kyushu Sangyo University, Cardiff University

Aaron Gibson
aaronlgibson@gmail.com
Kyushu Sangyo University

Luke Fryer
lukekutszikfryer@gmail.com
Kyushu Sangyo University

Abstract

Unlike classical test theory (CTT), where estimates of reliability are assumed to apply to all members of a population, item response theory provides a theoretical framework under which reliability can vary by test score. However, different IRT models can result in very different interpretations of reliability, as models that account for item quality (slopes) and probability of a correct guess significantly alter estimates. This is illustrated by fitting a TOEIC Bridge practice test to 1 (Rasch) and 3-parameter logistic models and comparing results. Under the Bayesian Information Criterion (BIC) the 3-parameter model provided superior fit. The implications of this are discussed.

Reliability under classical test theory

Test reliability refers to the internal consistency of a test. Ideally, a test taker who has not changed in language ability will receive similar scores on forms of a test regardless of how many times it is taken. Under Classical Test Theory (CTT), test reliability is most often measured by statistics such as Cronbach's Alpha and KR 20, but is perhaps most practically assessed using the Standard Error of Measurement (SEM). In simple terms, the SEM reflects the degree to which a test score may vary between test administrations by chance. Essentially a "standard deviation of error", it can be interpreted as a confidence interval for the learner's theoretical "true score". For example, if a student receives a scale score of 500 on a test and the test's SEM is 70, it can be said that if the test is retaken without improving in ability, the student can be expected to receive a score of 500 ±70 points 68% of the time. Setting an interval of two standard errors (±140 points) will yield a confidence interval of approximately 95% (for more information on the SEM, please refer to Brown, 1999).

While these statistics provide useful rules of thumb, it should be noted that under CTT, measures of reliability are assumed to apply to all test takers, regardless of the score they receive. Strictly speaking, this assumption is unrealistic. In situations where two test takers receive very high scores, it could be the case that the test questions used were not difficult enough to suitably challenge them, which could give us less confidence in score comparisons between them than we would have between two test takers of average ability, where the majority of questions were likely of suitable difficulty.

Another concern regarding how test reliability could change depending on student ability involves the use of multiple-choice questions. Suppose four students take a multiple-choice test of 100 questions, each with 4 options, and the test has an SEM of 5 points. One student receives a score of 65, one of 70, one of 20 and one of 25. Technically, the score difference between the first pair of students and second pair of students is identical at 5 points. But we have good reason to be less confident in the reliability of the scores for the second pair of students, because both received scores under the threshold of chance, as a score of 25 is possible simply by filling out answers at random. In cases where questions are too difficult to make a selection with any confidence, students become more likely to guess. Although multiple-choice formats likely affect reliability regardless of student ability, this practice could lead to more error for lower level students, where guessed answers could constitute a higher proportion of their total scores.

Reliability under item response theory

An advantage of Item Response Theory (IRT) is that under it, it is possible to examine how measures of test reliability change as a function of learner ability level (Embretson & Reise, 2000, p.185). We can determine not just how reliable a test will be overall, but how reliable it will be for particular groups of students with similar levels of language proficiency. This information can be used, for example, to determine if a test has suitable reliability at a proposed cut score between pass and fail, or if a new test is necessary for a special group of students of higher or lower ability than usual.

Under IRT, a number of factors can contribute to how reliable a test is considered to be for a given score (or, to use IRT terminology, the amount of "information" or precision the test provides). The following outline will focus on central concepts; Partchev (2004) provides details and formulas.

Item difficulty

Under the 1-parameter Rasch model, once misfitting items have been removed a primary determinant of reliability is how closely the difficulty of items used matches the ability level of the students tested. Items that students have a 0.5 probability of answering correctly are considered to provide the most information, and reciprocally to produce the smallest standard errors. The further items are in difficulty from a student's ability level, the less information the item will give about the student, and the greater standard errors of a student's ability estimate will become.

Item discrimination

In addition to item difficulty, the two-parameter logistic (2PL) model uses item discrimination to calculate the information items provide about student ability. Although items are still believed to provide maximum information when students have a 0.5 probability of answering them correctly, if difficulties are equal, items with low discrimination are considered to provide less information than items with high discrimination.

Guessing

In addition to item difficulty and discrimination, the three-parameter logistic (3PL) model also uses the likelihood a student of very low ability will choose a correct answer to determine item information. If a student's probability of correctly answering a question is as low as the probability of correctly guessing it by chance, the item is not considered to provide any information about student ability. An interesting aspect of the 3PL model is that rather than providing maximum information for students with 0.5 odds of correctly answering, maximum information falls at the midpoint between the odds of a correct guess by a very low level student and 1. For example, if an item has a 0.2 probability of being answered correctly even by a very low level student, the item is considered to provide maximum information for students who have a 0.6 probability of answering it. The practical result of this is that unlike under CTT and Rasch frameworks, tests that result in mean scores of 50% are not always optimal; due to the fact that some questions will be answered correctly by chance, the ideal mean score can be somewhat higher.

Given these different considerations, estimates of test information and resulting reliability can vary greatly depending on the item response model used. Although person ability and item difficulty estimates are typically very highly correlated regardless of the IRT model used (Stewart, 2012), the same cannot be said about test information and resultant reliability, which can vary substantially between models (De Ayala, 2009). Model fit must be examined to determine which model best describes a test.

Aims

In this paper we will examine a practice form of a well-known and widely used test of English language proficiency, the TOEIC Bridge test (ETS, 2010) under CTT and 1- and 3-parameter item response models, in order to demonstrate how estimates of reliability differ under each framework. We will then conduct a model fit comparison to determine which IRT model is most appropriate for the data, detail how test reliability varies by student ability level under the chosen model, and explain the significance of the findings in practical terms.

The TOEIC Bridge test

The TOEIC Bridge (ETS, 2010a) is a test of emerging English language ability for learners with proficiency too low to be measured by the better-known TOEIC Test. Items used are similar in format and content to those of the TOEIC Test, but of lesser difficulty. Like the TOEIC Test, it has Listening and Reading sections, though with only 50 items each and 100 total, as opposed to 100 items each for 200 total. Although it is not yet popular with many private test takers, it is increasingly used by institutions such as junior high schools, high schools and low-level universities; of the approximately 198,000 people who took the test in 2009, 98% took the TOEIC Bridge IP (ETS, 2010b), which is delivered to such institutions and administered on-site rather than at a testing ground operated by ETS. Scores are derived from raw scores on the test, without penalty for guessing. Scores for its two sections are converted to scale scores ranging from 10-90. Total scale scores for both sections range from 20-180 (ETS, 2007a). In 2009, the mean total scale score for TOEIC Bridge IP Test (which constitutes the vast majority of tests taken) was 118.1 (ETS, 2010b).

In order to prepare for an official administration of the TOEIC Bridge IP used as an achievement test, 1071 first and second year students at a private university took an official ETS TOEIC Bridge practice test included in a test preparation workbook (ETS, 2008a). The mean scale score for the students at the private university on the official test, written shortly after, was approximately 120, close to the average reported by ETS for all TOEIC Bridge IP test takers the previous year (ETS, 2010b).

Initial results

The practice test had a KR-20 reliability statistic of 0.85, and a correlation of 0.81 to scores on the official test, which students took shortly after. The test's raw score mean was 50.14, with a high raw score of 86 and a low score of 22. The standard deviation was 11.72. As per Brown (1999), this results in a standard error of measurement of approximately 4.5 points.

The mean item difficulty was 0.5, which, as with the Rasch and 2PL IRT models, is considered optimal for norm-referenced tests under CTT (Brown, 2005), as such items aid in producing a normal distribution of scores. This results in a mean score near 50%, which, being the midpoint of possible scores, is typically considered ideal for norm-referenced tests under CTT. As the mean TOEIC Bridge score of test takers was close to the average score for all TOEIC Bridge IP test takers as reported by ETS, indicating that the sample's ability level was close to the mean of the average TOEIC Bridge IP test taker, it is possible that this is by design. However, the dispersion of scores of the current sample was quite narrow, with a standard deviation of 11.72, and two standard deviations out yielding a score range between roughly 27 and 73. As this score range covers almost 98% of the tested population, only approximately 50% of the potential scores are applicable to the majority of learners tested.

Table 7. Mean item difficulty by test section and part

Test Section	Part	Mean	K	SD	Min.	Max.
Listening	I	0.63	15	0.26	.24	.97
	II	0.55	20	0.20	.27	.96
	III	0.41	15	0.11	.20	.59
	Total	0.53	50	0.21	.20	.97
Reading	IV	0.45	30	0.15	.21	.87
	V	0.50	20	0.23	.20	.89
	Total	0.47	50	0.18	.20	.89
Total Test		0.50	100	0.20	.20	.97

Examining test reliability under Rasch and 3PL IRT models

The test data was analyzed under 1- and 3-parameter logistic item response models using JMP 8. Item parameter estimates were then used to calculate test information functions (IRT reliability estimates) for both models in Microsoft Excel, using formulas detailed by Parchev (2004), and graphed using guidelines by Kim (2004). The "Test Information" on the vertical axis refers to the precision of the test for given levels of student ability, listed on the horizontal axis. An ability (or "theta") level of 0 indicates the student average, as the theta mean is person-centered. The greater the test information at a given ability level, the more reliable the test is considered to be at that point.

The test's test information function (TIF) differed markedly between models. Under the Rasch model, the TIF appears to be ideal, with maximum information provided for test takers of average ability of 0, corresponding to the mean score of nearly 50% under classical test theory analysis; under both the Rasch model and CTT, test reliability is optimal for the majority of students.

However, under the 3PL model, which estimates the probability that a test taker of very low ability will correctly guess an answer by chance, the TIF is considerably more uneven, with maximum information given for test takers with ability estimates between approximately 1-2.5 logits. Much of the discrepancy can be accounted for by the 3PL model's use of the likelihood of a very low level student guessing the answer by chance in its estimate of reliability. If the model is accepted, the implication is that a somewhat easier version of the test would have maximum reliability for the majority of the students tested in this study.

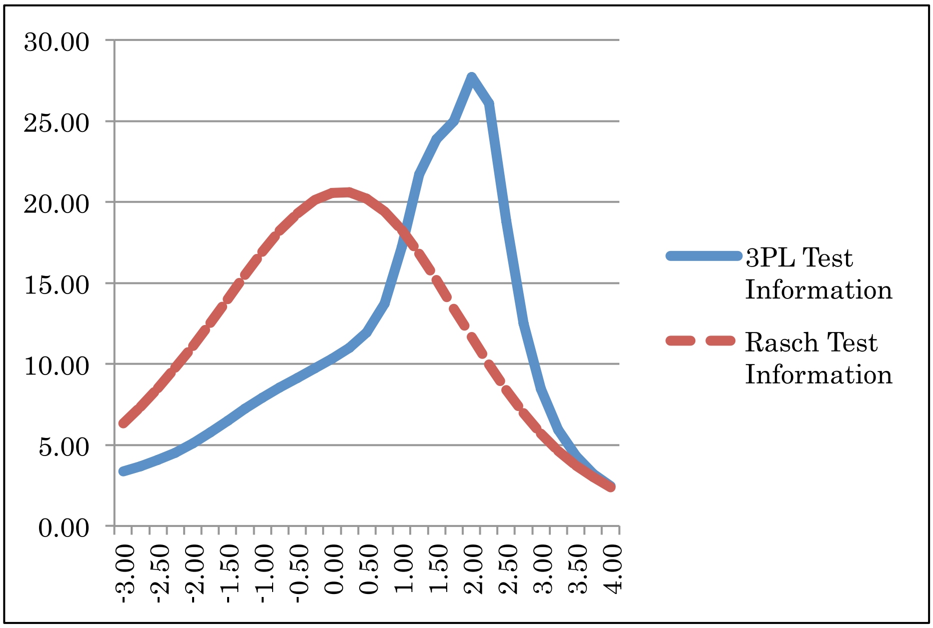

***Figure 1.* Test information functions for Rasch and 3PL models**

Model selection

The two models tell us different stories about the test's reliability. How can we tell which is closer to the truth? To do this, we must make model fit comparisons, which can be performed in the statistical software program R (R Core Development Team, 2008), using the IRT package LTM (Rizpolous, 2006). In this issue, software columnist Aaron Batty explains how to get started with R using the LTM-compatible graphical user interface RKWard (though this particular feature of LTM must still be requested by command line).

When assessing model fit, it is important to consider that, to at least a negligible degree, nested models with more parameters (such as the 3PL when compared to the 1PL) will nearly always demonstrate superior fit, but that these solutions may simply represent an overfit of the model to the data set, and a solution that will not necessarily be generalizable to other samples (Zucchini, 2000). However, an earlier analysis by the authors on the practice test data set demonstrated that under several criteria, including the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which penalize more complex models, the 3PL model displayed superior fit to the 1PL Rasch model. Therefore, the 3-parameter model can be considered to give a more accurate description of the test form's properties.

Table 8. Likelihood ratio table for Rasch and 3PL models

Model	AIC	BIC	Log Lik.	LRT	df	p
Rasch	114105.6	114558.3	-56961.81			
3PL	112644.2	114002.2	-56049.10	1825.43	182	<0.001

Examining test reliability under the selected model

Test information for different levels of student ability can be used to calculate standard errors (see Partchev, 2004), which are shown in the table below. Student ability under the 3PL model was equated to TOEIC Bridge Scale Scores, albeit crudely, using equipercentile equating. It should be noted these are rough estimates, as only a subsample of 491 for which the scores of both tests were available was examined, and score distributions were not smoothed. For a primer on more sophisticated methods of equipercentile equating, please refer to Livingston's eminently readable guide on the subject (2004), or Kolen & Brennan's authoritative book on test equating (2004).

Table 9. Standard errors and estimated scale scores for 3PL model.

Estimated TOEIC Bridge Scale Score	82	84	88	96	102	110	118	128	136	142	150	154
3PL Theta	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5
Theta Standard Error	0.54	0.50	0.44	0.39	0.36	0.33	0.31	0.29	0.24	0.20	0.19	0.23
Test Information	3.4	4.1	5.1	6.5	7.9	9.1	10.3	12.0	17.3	23.9	27.7	18.8

The results suggest that standard errors are lower for students with ability levels between 1-2 logits than they are for students of mean ability (0). Under the 3PL framework, then, the test may be most reliable for test takers with scale scores between roughly 130 and 150. Due to the limitations of the smaller examined sample, it is not possible to equate higher scale scores with this data set, although under the 3PL model, standard errors should increase for test takers of higher ability.

Practical implications

An implication of these results is that a somewhat easier version of the test may result in higher test reliability for the majority of students, as the item difficulty would closer match the sample's mean ability level. To relate these findings to raw scores, as the TOEIC Bridge test is multiple-choice, it is exceedingly difficult for a student who answers every question to receive a score of less than 20%, even if they do not actually know any of the answers. Consequently, a score of 0 does not represent the "true" minimum score of the test, meaning a score of 50% does not represent the midpoint between minimum and maximum scores, or an ideal point to center the score distribution. Instead, items that result in a mean raw score of roughly 60% may be closer to ideal.

Taking this information into account, we have found that when we make our own norm-referenced multiple-choice tests for this student population, even if IRT models are subsequently ignored in favor raw scores and classical analyses, tests with means of 60% do, in fact, appear to result in higher reliability than tests constructed from the same item bank with means of 50% when other item properties are equal. Due to such experiences, we have found that although the large sample sizes required for estimation of some IRT models (Over 1000 for the 3PL, for example) can make such studies troublesome to conduct, analyses of reliability under IRT can result in meaningful improvements to language tests.

References

Brown, J. D. (1999). Standard error vs. Standard error of measurement. *Shiken, 3*(1) p.20-25

De Ayala, R. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

ETS. (2007a). *TOEIC® Bridge Examinee Handbook.* Retrieved September 10, 2010, from http://www.ets.org/Media/Tests/TOEIC_Bridge/pdf/TOEIC_BridgeExam.pdf

ETS. (2008a). *TOEIC Bridge kōshiki wākubukku*. Tokyo: Kokusai Bijinesu Komyunikēshon Kyōkai TOEIC Un'ei Iinkai.

ETS. (2010a). *About the TOEIC Bridge.* Retrieved September 10, 2010, from http://www.toeic.or.jp/toeic_en/bridge/about.html

ETS. (2010b). *TOEIC Bridge® Data & Analysis 2009.* Retrieved 10 September 2010, from http://www.toeic.or.jp/bridge/pdf/data/Bridge2009_DAA.pdf

Kim, J (2004) An Excel manual for item response theory. Retrieved 20 August 2012, from http://education.gsu.edu/coshima/EPRS8410/Sarah_Project1%2012%203%202004.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York, NY:Springer.

Partchev, I. (2004). A visual guide to item response theory. Retrieved 18 August 2012, from www.metheval.uni-jena.de/irt/VisualIRT.pdf

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rizpolous, D. (2006). ltm: An R package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(17), 1-25.

Stewart, J. (2012) Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination. *Shiken Research Bulletin, 16*(1). 15-22.

Zucchini, W. (2000). An Introduction to Model Selection. *Journal of Mathematical Psychology, 44*(1), 41-61. doi:10.1006/jmps.1999.1276