Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination

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

Abstract

It has been frequently stated that Rasch and Item Response Theory produce interval-scale measures where raw scores can only provide ordinal measures, and that therefore, researchers should choose Rasch/IRT measures when selecting variables for common statistical tests (Wright, 1992; Harwell & Gattie, 2001). In this study, this claim is empirically examined by conducting Pearson Correlations and ANOVAs on two data sets using raw scores, Rasch Person Measures and 2-Parameter IRT ability estimates, in order to determine if results differed as a consequence. Raw Scores and Rasch Person Measures were very highly correlated, and lead to extremely similar results in all cases. For a well-constructed, reliable test the same was true of 2PL ability estimates. However, in cases where the test has middling to poor reliability, 2PL ability estimates appear to produce a somewhat more sensitive measure of a latent trait than raw scores, which can result in meaningful differences in statistical tests.

Introduction

Noteworthy proponents of Rasch Measurement and Item Response Theory (IRT) have argued that raw test scores used under Classical Test Theory (CTT) are ordinal and non-linear in nature, and therefore not suitable for use in "normal" (i.e., parametric) statistics (e.g. Wright, 1992). The theoretical argument underlying this claim has led proponents of Rasch Measurement and Item Response Theory to warn against the use of raw scores from psychological tests as variables in experiments and statistical analyses. As Harwell and Gattie (2001) wrote in Review of Educational Measurement, "educational researchers frequently employ ordinal-scaled dependent variables in statistical procedures that assume that these variables possess an interval scale of measurement . . . . data possessing an ordinal scale will not satisfy the assumption of normality needed in many statistical procedures and may produce biased statistical results that threaten the validity of inference" (p. 105). The conversion of test data to raw scores was illustrated using the Rasch model.

Harwell and Gattie explain how to rescale the theoretically ordinal data produced by raw scores into theoretically interval data using item response theory, but they do not illustrate the usefulness of the rescaled data in a common experimental design, for example, comparing the use of Rasch measures in a t-test or ANOVA rather than raw scores. When using scores on a psychometric test as a dependent variable, does the substitution of raw scores for Rasch/IRT Measures have a practical effect on a statistical analysis?

Although theoretical arguments for the superiority of Rasch and Item Response Theory over Classical Test Theory (CTT) abound, there has been a relative paucity of empirical research regarding the practical benefits of using ability estimates calculated under item response theory over raw scores in this regard. Noting this, Fan (1998) argued, "Theoretical models are important

in guiding our research and practice. But the merits of a theoretical model should ultimately be validated through rigorous empirical scrutiny" (p. 15). In Fan's study comparing CTT, Rasch, 2PL IRT and 3PL IRT estimates of item difficulty and person ability, he found that all measures were highly comparable and very highly correlated. Though he acknowledged the usefulness and practical advantages of Rasch and IRT for applications such as item banking and computer adaptive testing, he concluded that differences between CTT and forms of IRT were overstated, and that despite theoretical arguments to the contrary, the actual differences between the two approaches were minimal. Various subsequent studies (e.g., MacDonald & Paunonen, 2002; Progar, Socan & Pec, 2008) have essentially confirmed these findings.

Of course, high Pearson correlations do not necessarily indicate that IRT estimates and raw scores are identical in every respect. As Linacre (1998) explains, even if two measures are perfectly correlated, the length of intervals between scores can vary greatly. Since these intervals are accounted for by logits, change scores of the same numerical value can be considered equivalent, though the same often cannot be said of raw scores (Embretson & Reise, 2000). Still, the strong similarities are discouraging, and do little to encourage researchers outside of psychometrics to move from raw scores to modern test theory approaches.

Suppose a researcher in SLA wishes to compare the effects of two teaching approaches on test scores. Theoretical arguments aside, and even supposing a given IRT model is shown to have better fit to the data using a statistic such as the BIC, ultimately do the approaches in scoring methods genuinely differ enough in practice that one should be preferred over others? The question is of central importance to any language acquisition researcher who has considered using Rasch or IRT measures for tests or surveys, because a common assumption is that the use of such models will improve the measurement properties of the research instruments they are applied to, and that by extension the likelihood of detecting a statistically significant difference between treatments in a common statistical analysis will increase. If this is not the case, it is more difficult for researchers outside of psychometrics and educational assessment to justify the time and expense of adapting to modern test theory and buying and learning to operate the software programs required to operationalize it.

Research Questions

Consequently, this short paper will attempt to answer three questions:

If used in place of raw scores, do Rasch/2PL IRT ability estimates improve the correlation of one language test to another?

Are the results of an ANOVA noticeably different if Rasch/2PL IRT ability estimates are used as the dependent variable rather than raw scores?

Do results differ if the above experiments are conducted using less reliable tests?

Method and Analyses

Data from two tests were used in this study, one with a high reliability (a Cronbach Alpha of 0.91) and one with a relatively low reliability (a Cronbach Alpha of 0.75). This was done to test if Rasch and IRT ability estimates were effective in reducing measurement error in a less reliable test. The tests' descriptive statistics are listed below, in Table 1. In addition to responses, Data Set A also included test takers' scale scores on a second test, the TOEIC Bridge, and information regarding the test takers' teachers for Listening classes. Data set B included information on the

individual classes each student was registered for. This additional information allowed for a Pearson correlation from scores to a second test and ANOVAs with categorical independent variables. It should be stressed that these analyses are essentially ad hoc, and done merely to examine differences between raw scores and IRT ability estimates in common statistical tests. Therefore, little attention will be given to the significance or meaning of the various results.

Table 5. Descriptive statistists of tests used.

	A: 2010 KSU Test	B: Western Music Test
k	100	100
N	654	234
Mean Score	48.50	56.40
SD	16.60	8.00
Reliability	0.91	0.75

Analyses using the 2010 KSU Test

The 2010 KSU Test is an older form of a placement test of English language listening and reading skills currently under piloting at the author's institution. In addition to test scores, the data set includes the TOEIC Bridge test scores of the students who took it, and information regarding each student's teacher. This permitted a) a Pearson correlation between the test and an external, validated measure of language proficiency, the TOEIC Bridge test, and b) an ANOVA of the effect of students' teachers on test scores.

In addition to raw scores on the test, ability estimates for the test under the Rasch model and the 2-Parameter Logistic Model were generated using the statistical software package JMP 8. A drawback of this study is that the JMP manual does not specify its estimation method, but it was observed that ability estimates produced for the Rasch model were identical to those produced by the program WINSTEPS, which uses JMLE.

The Pearson Correlation and ANOVA described above were then performed three times, each time using a different ability measure of the test as a variable: raw score, Rasch person measure, 2PL ability estimate. Analyses were then compared to determine if use of Rasch and 2PL ability estimates substantially altered the results of the experiments. Correlations between the raw scores of the test and Rasch and 2PL ability estimates are listed below. As Fan reported, correlations are very high, though very marginally lower between raw score and 2PL ability estimates.

Table 6. Correlations of raw scores of 2010 KSU Test to Rasch and 2PL IRT ability estimates.

	Raw Score
Rasch	0.997
2PL	0.986

Next, the test was correlated to the TOEIC Bridge test using the three scoring methods, as listed below in Table 3. The correlation for each is approximately 0.83; the difference between them is essentially indistinguishable.

Table 7. Correlations of 2010 KSU Test to the TOEIC Bridge Test by raw score and Rasch and 2PL IRT ability estimates

	TOEIC Bridge Test
Raw Score	0.833
Rasch	0.834
2PL	0.835

Next, an ANOVA was conducted on the effect of students' Listening class teachers on their scores, using the test's listening section scores calculated under all three methods. The F-Test was significant for each treatment (p = < .0001), though the R-Square and Adjusted R-Square for raw scores was slightly higher (marked in bold) than either IRT scoring method.

Table 8. One-way ANOVAs of differences on KSU Test scores by teacher using raw scores, Rasch measures, and 2PL ability estimates as dependent variables.

	Dependent: Raw Score	Dependent: Rasch Person Measure	Dependent: 2PL Ability Estimate
F Ratio	15.603	14.638	15.079
Prob > F	< .0001*	< .0001*	< .0001*
R-square	**0.241**	0.229	0.234
Adj R-square	**0.225**	0.214	0.219
Root Mean Square Error	7.396	0.772	0.884
Mean of Response	25.199	0.000 (person centered)	0.000 (person centered)
Observations (or Sum Wgts)	654	654	654

Analyses using the Western Music Test.

The Western Music Test was a less successful form of a low-stakes classroom test of students' listening comprehension of songs studied throughout a semester, and understanding of the expressions and vocabulary found in the lyrics. Reliability was low due to poorly targeted items and a low average point biserial correlation of approximately 0.19. Despite poor overall reliability no item point-biserial correlations were substantially negative, though two were effectively 0. Removing them did not appear to significantly improve split-half reliability.

The use of a test of lower reliability allows further examination of the theoretical advantages of ability estimates under the 2-parameter model. Although it does not pertain to the argument that IRT measures produce interval data, a potential advantage of the 2PL ability estimates is its use of item slopes in calculating ability. Under the Rasch model, items are assumed to have equal or approximately equal discrimination, meaning that equal weight is given to each question answered correctly. In contrast, the 2PL model weighs each item by its item discrimination, meaning successful endorsement of items with high discrimination contributes more to estimates of ability than items with low discrimination.

However, were the items of a test to fit the Rasch model and have roughly equal discrimination, any advantages offered by the 2PL ability estimate would become unobservable. This could be the case with the 2010 KSU Test, which was constructed in accordance to its ideal Test Information Function under the Rasch model, and uses items of fairly high (and relatively equal) discrimination for a multiple-choice test. Were 2PL ability estimates able to reduce measurement error, such an effect would likely be most measurable on a test with higher degrees of measurement error to begin with. A drawback of Fan's study was that descriptive statistics were not reported for the test analyzed, the Texas Assessment of Academic Skills (TAAS) for the 11th grade. Presumably, however, the State of Texas School Board employs tests of high reliability. If so, it could be argued that Fan's data set was not ideal for testing this possible advantage of 2PL ability estimates.

The test's raw scores and ability estimates under the Rasch and 2PL models were correlated, as can be seen in Table 5.

Table 9. Correlations between raw scores, Rasch person measures, and 2PL ability estimates of a test with a Cronbach alpha of 0.76.

	Raw Score	Rasch	2PL
Raw Score	1.000		
Rasch Person Measures	0.999	1.000	
2PL Ability Estimate	**0.935**	**0.942**	1.000

Correlations between raw scores and Rasch measures are very nearly 1. In this instance, however, the correlation between 2PL ability and raw scores is noticeably lower. What effect could this difference have on an experiment conducted using test scores as a variable? Although there is no data for a second test with which to correlate to, we can hypothesize how the test could correlate to another test of equivalent reliability under each ability measure by examining split-half correlations under each scoring method, as can be seen in Table 6.

Table 10. Split-half correlations for raw score and Rasch and 2PL ability estimates.

Raw Score	0.504
Rasch Ability Estimate	0.506
2PL	**0.571**

In this case, the difference in correlation is fairly sizeable, and would be enough to warrant the use of 2PL ability estimates for comparisons of test scores using Pearson correlation.

Finally, an ANOVA was conducted on test scores by class using each ability estimate method. Although differences in R-squared values were small, in this case there was a critical distinction: the ANOVA using the 2PL ability estimate had a p-value well below the critical threshold of 0.05,

and the other two ability estimates did not. In this instance, a technically statistically significant result was reached that would not have been had raw scores been used. Though the substantive difference remains negligible and <0.05 is a rather arbitrary value for "significance" (See Eidswick, this issue and Brown, this issue for further discussion), regrettably at many journals it can still make the difference between results that are considered publishable and results that are not.

Table 11. One-way ANOVAs of differences on test scores by class using raw scores, Rasch measures and 2PL ability estimates as dependent variables.

	Dependent: Raw Score	Dependent: Rasch Person Measure	Dependent: 2PL Ability Estimate
F Ratio	1.838	1.854	2.301
Prob > F	0.072	0.069	**0.022***
R-square	0.065	0.066	0.080
Adj R-square	0.030	0.030	0.045
Root Mean Square Error	8.651	0.466	0.978
Mean of Response	56.355	-0.003	-0.015
Observations (or Sum Wgts)	220	220	220

It appears that although differences are negligible for more reliable tests, for less reliable tests 2PL ability estimates can have greater efficacy than either raw scores or Rasch measures in statistical tests. To my knowledge this has not been documented in the literature. When told of it, experts in Item Response Theory have expressed surprise; Hambleton (personal communication) doubted that there could be much difference unless item point-biserials were negative. However, while this research marks my first formal study of the phenomenon, I have observed it in the past with a fair amount of consistency. It could be the case that researchers in psychometrics and educational assessment have a tendency to work with highly reliable tests that are less likely to reveal such differences to begin with. Unfortunately, however, testing instruments of poor reliability can still be quite common in other fields, and it is still possible for research using tests and surveys with reliability as low as 0.75 to be published in second language acquisition journals. Borsboom (2006) lamented that researchers in the social sciences often ignore advances in psychometrics and modern test theory. In a reply, Kane (2006) stated, "A great way to get their attention is to show them what you can do for them." Perhaps this application of 2PL ability estimates provides such an example.

Conclusions

In conclusion, raw scores and Rasch ability estimates are very highly correlated, and lead to extremely similar results when used in common statistical tests. For a well-constructed, reliable test, the same is true of 2PL ability estimates. However, in cases where the test has middling to poor reliability, 2PL ability estimates actually do appear to produce a somewhat more sensitive measure of a latent trait than raw scores, and their use as variables can result in meaningful differences in statistical tests.

I do not mean to diminish the usefulness of Rasch measurement, however. In closing, three things must be remembered:

1. *Rasch analysis should be seen as a method for examining and altering tests to reflect optimal measurement properties, not a magic transformation that automatically improves data.*

It should be noted that the first test examined had already been optimized in Winsteps, with items chosen to produce the optimal Test Information Function under the Rasch model. Misfitting items from pilot stages were not used in the final form. Therefore, although the final form worked just as well with raw scores as with Rasch measures, it had already benefitted from analysis under a Rasch framework; the analysis can be of benefit even if the raw scores of the resulting test are used.

2. *Rasch measures are of practical value even with near-perfect correlation to raw scores.*

Rasch measures provide a theoretical framework with which to examine interactions between test takers and individual items. They remain useful in equating scores between separate test forms, where different raw scores can result in identical measures. They can then be used as the basis for scale scores between forms. There is furthermore a persuasive theoretical argument that by accounting for intervals of difficulty between items, Rasch measures allow for comparable change scores, ensuring that, for example, the difference between the scores of 55 and 59 is indeed identical to the difference between the scores of 90 and 94.

3. *The 2PL ability estimates are only superior to raw scores if the test itself is inferior.*

While the 2PL ability estimates may be of value in salvaging data when a research instrument proves to have less than ideal reliability (and for whatever reason re-doing the experiment is no longer possible), it should be stressed that this technique should be viewed as a distant second to simply constructing a reliable test in the first place. As can be seen with the first data set, if the test is well constructed, results by all three scoring methods will be essentially identical.

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440, DOI: 10.1007/s11336-006-1447-6

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381. DOI: 10.1177/0013164498058003001

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131. DOI: 10.3102/00346543071001105

Linacre J.M. (1998). Do correlations prove scores linear? *Rasch Measurement Transactions, 12*(1), pp. 605-606. Retrieved January 16, 2012 from http://www.rasch.org/rmt/rmt121b.htm

MacDonald, P., & Paunonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62.*

Rupp, A. & Zumbo, B. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64*(4), 588-599. DOI: 10.1177/0013164403261051

Wright, B. D. (1992). Raw scores are not linear measures: Rasch vs. Classical Test Theory CTT Comparison. *Rasch Measurement Transactions, 6*(1), 208. Retrieved January 16, 2012 from http://www.rasch.org/rmt/rmt61n.htm