RASCHING AN ACHIEVEMENT TEST

Larry R. Nelson

Burapha University, Chonburi, Thailand

Curtin University of Technology, Western Australia

DRAFT 27 May 08 of: An invited paper prepared for Volume 6 of the
Thai Journal of Educational Research and Measurement,
ISSN 1685-6740, to be published in 2008.

In this paper I relate experiences gained in the process of moving data from Lertap, a classical test theory (CTT) system, to ConQuest and Winsteps, two item response theory (IRT) systems which have a particular focus on the Rasch model. The work reported below involves matters of dimensionality, indices of model fit, measurement precision, and scaling.

To a considerable extent this research was prompted by comments brought back by the Dean of a large engineering faculty after she had attended a recent conference. "We might have a look at Rasch scaling", she wrote on a memo distributed to staff. The memo quoted from a conference paper: "Test scores are mere enumeration, a tally of number of items correct. They are not measures of proficiency. Rasch scales are objective fundamental measures expressed on an interval scale".

Of course test scores have a bit more going for them than simple enumeration. If on a sixty-item test, scored on a number-correct basis, Jorge has a score of 60, Milagros a score of 45, and Jaime a score of 30, we can say that Jorge has answered twice as many items correctly as Jaime. If we imagine a horizontal scale, with number-correct score marked out on it in equally-spaced intervals, the distance between Jaime and Milagros will be the same as that between Milagros and Jorge, and the meaning will be the same: 15 more items answered correctly.

Our scale has counts of items correct; on this scale a change of one unit in the positive direction indicates one more test item correct. Scale intervals have equal spacing, and the scale even has a meaningful zero point. A student with a test score of zero is a student who did not answer a single item correctly. We have a ratio scale indicating number of test items answered correctly.

The problem is that we cannot conclude that the student with a score of zero knows nothing about the content covered by the test. The test may have been too hard for the student; given easier items he or she may well have a better score. While a direct relationship between number of items correct and subject dominance is assumed, and natural, the items are but a sample, generally a very small one, of the content universe.

Lack of a true zero point makes it impossible to conclude that Jorge knows twice as much as Jaime. We cannot even say (and this is the nexus of the situation) that the distance between Jaime and Milagros has the same meaning as the distance between Milagros and Jorge. Such distances only relate to the number of items correct.

What we would like to do is be able to plot students on a scale where units measure "ability", "subject mastery" or "proficiency". The zero on this scale will indicate a true

absence of whatever it is we are measuring, and a move to the right on the scale will take us over more of whatever it is we are measuring, with a change of one unit anywhere on the scale having the same meaning.

Can we achieve these objectives by turning to Rasch scaling? No, not all of them. Rasch aspires only to produce interval measures; it will not give us a ratio scale where we know what zero means. Still, the Dean's memo suggests that having interval measures might be a worthy advance. "Why don't we give Rasch a go to see what it produces", she wrote.

Not everyone agrees that Rasch delivers. The Rasch model is one which requires all test items to have equal discrimination. Some find this undesirable. Bock (in du Toit, 2003, p.840) writing about Rasch: "*Although this solution to the item-parameter estimation problem is of interest theoretically, it does not satisfy the requirements of practical testing programs …. It also assumes the item slopes to be equal when more often than not in practical testing they are unequal …. As a result, there is no possibility of estimating item discriminating powers, which are essential in test construction for choosing items that ensure good test reliability and favourable score distributions*".

In this study I take an achievement test from the engineering faculty and look at it from both CTT and Rasch perspectives. The test in question is one whose items exhibit a range of difficulty and discrimination, making it, in my experience, typical of its kind. Will test results fit the Rasch model? If they do, how can I subsequently benefit from the linear Rasch scores, the "measures" which are a primary objective? Can I be sure that I do indeed have interval measures from Rasch?

The data set

The achievement test used in this study was developed by staff at the engineering faculty. It has been used for numerous years and is regularly subject to quality control revision.

Sixty (60) multiple-choice items are involved in the test. Each item has four options and is scored on a right-wrong basis. The items are designed to test over knowledge, comprehension, and application levels, with an emphasis on application.

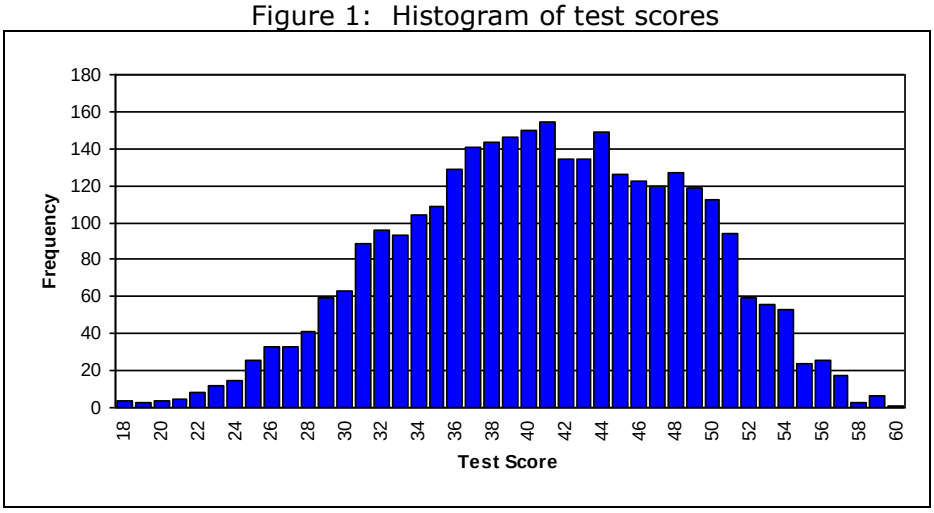
Test results are used to class students as fail, pass, credit, distinction, and high distinction. In a typical year, from five to ten percent of students will achieve at the high-distinction level. This is regarded as a high-stakes exam; students must pass it in order to meet one of the requirements of the professional engineering qualification they are seeking.

The test results which form the basis of the present paper were collected recently. Over three thousand students sat the test in one of several campus test venues, all professionally monitored and controlled.

The answers collected from the test venues were processed using Lertap (Nelson, 2000). An initial run with all results showed a negatively-skewed score distribution with a median of 41, mean of 40.80, and standard deviation of 7.75. With one exception, scores started at 18 (four students), and continued up to the maximum possible score of 60 (one student). The exception was a single student with a score of 2, a student who answered only the first twelve items, getting two of these correct, and failing to even guess at the remaining forty-eight items. This student's results were deleted from the dataset. The mean of the test scores was then 41.86, standard deviation 7.72.

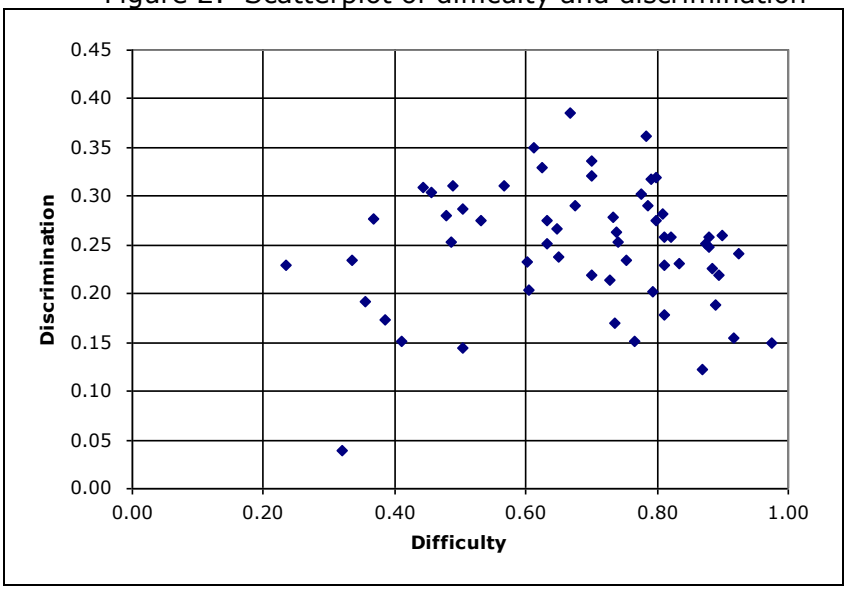
The distribution of test scores is indicated in Figure 1. Note that this figure does not begin at zero, the lowest possible score. If it did, it would be more apparent that the test was on the easy side, with the central cluster of scores around 41, corresponding to a percentage score of 68%.

Figure 1: Histogram of test scores



A scatterplot of item difficulty by discrimination is shown in Figure 2, where the discrimination index is a point-biserial correlation coefficient corrected for part-whole inflation.

Figure 2: Scatterplot of difficulty and discrimination



Sixteen of the sixty test items had difficulties in the 0.80 to 0.89 range; five had difficulties at or above 0.90. Item 42 was the easiest item, with a difficulty of 0.97; only a smattering of items tended to be hard for the students (had a difficulty index below 0.40).

Item 8 had the lowest point-biserial in the total group (0.04); item 47 had the highest point-biserial (0.39), followed by item 50 (0.36).

Given the distribution of item difficulty values seen in Figure 1, the average test score of 40.82, just under 70%, might be expected. If this was the minimum score required to

at least pass the test, 49% of the students would get over the line. A test score of 53 or greater (88%; z= 1.572) would put a student into the upper 5.9% of the scores, and might qualify as representing high distinction; 186 students got a score at or above this level.

Lertap found the test to have a reliability of 0.83, as measured by coefficient alpha. The corresponding classical standard error of measurement was 3.21. In the area of the average score, Lertap reported a conditional standard error of measurement value of 3.33, while in the vicinity of a (true) test score of 53, the high-distinction cutoff, Lertap reported a conditional standard error of measurement 2.30. These conditional values were computed by using a compound binomial model suggested by Lord (1984), and often referred to as "Lord's Method IV" (also see Nelson, 2007).

Were we to transform the raw test scores to a scale with mean 500, standard deviation 100, the classical standard error of measurement on the new scale would be 41.63, and the conditional standard error of measurement at the mean, 500, would be 43.17. For a raw test score of 53, corresponding to a scaled score of 657, the conditional standard error of measurement would be 29.79. These standard error values will later be compared to corresponding figures derived from Rasch scaling.

Dimensionality

It is well known that IRT methods depend on unidimensional items; having but a single common factor is critical to IRT.

I looked at the question of dimensionality using a scree test and eigenvalue comparisons, following Nelson (2005). For the sake of brevity I will not show the scree plot here; Table 1 represents the eigenvalues of the correlation matrix with ones on the diagonal (a principal components resolution).

Table 1: Eigenvalue statistics

EV Number	1	2	3	4	5	6	7	8	9	10
Magnitude	5.67	1.78	1.43	1.25	1.24	1.18	1.16	1.13	1.10	1.09
Variance %	9.4%	3.0%	2.4%	2.1%	2.1%	2.0%	1.9%	1.9%	1.8%	1.8%
R^2	.24	.83	.94	.98	.98					

Another approach to scree plotting is suggested by the IRT Modeling Lab of the Department of Psychology, University of Illinois (io.psych.uiuc.edu/irt/): when items are scored on a right-wrong basis, they recommend the use of tetrachoric coefficients in the inter-item correlation matrix. Lertap was directed to do this, producing the results seen in Table 2.

Table 2: Eigenvalues from tetrachoric matrix

EV Number	1	2	3	4	5	6	7	8	9	10
Magnitude	9.74	2.40	1.78	1.53	1.42	1.31	1.29	1.22	1.19	1.16
Variance %	16.2%	4.0%	3.0%	2.6%	2.4%	2.2%	2.2%	2.0%	2.0%	1.9%
R^2	.23	.83	.94	.97	.98					

Has the use of tetrachoric correlations improved my ability to answer the dimensionality question? Yes, it seems so. In Table 1, the difference between the first and second eigenvalue is about 3.9. The ratio of these two values is 3.2; the ratio of the second eigenvalue to the third is 1.3, while the ratio of the third to the fourth eigenvalue is 1.1.

In Table 2, the difference between the first and second eigenvalue is greater, 7.3, with a ratio of just over 4.0 between these two eigenvalues. The respective ratios of the 2nd/3rd and 3rd/4th eigenvalues are 1.3 and 1.2.

We could also consider the variance accounted for measure, seen as Variance % in the two tables – that for the leading eigenvalue is much higher in Table 2.

These results would add support for suggesting that there is one dominant factor, or dimension, underlying the sixty test items. (For more on the interpretation of eigen-values and scree tests, see Nelson 2005, and the IRT Modeling Lab at www.io.psych.uiuc.edu/irt/.)

As it turns out, looking at dimensionality before going to Rasch may not be necessary. I did not realise this as I started my work; when I subsequently began to apply the Winsteps program it became apparent that Winsteps has its own principal components resolution of residuals, and another sort of scree plot. This is definitely an alternative to looking at dimensionality beforehand. At least, this is what I thought at first. More on this below.

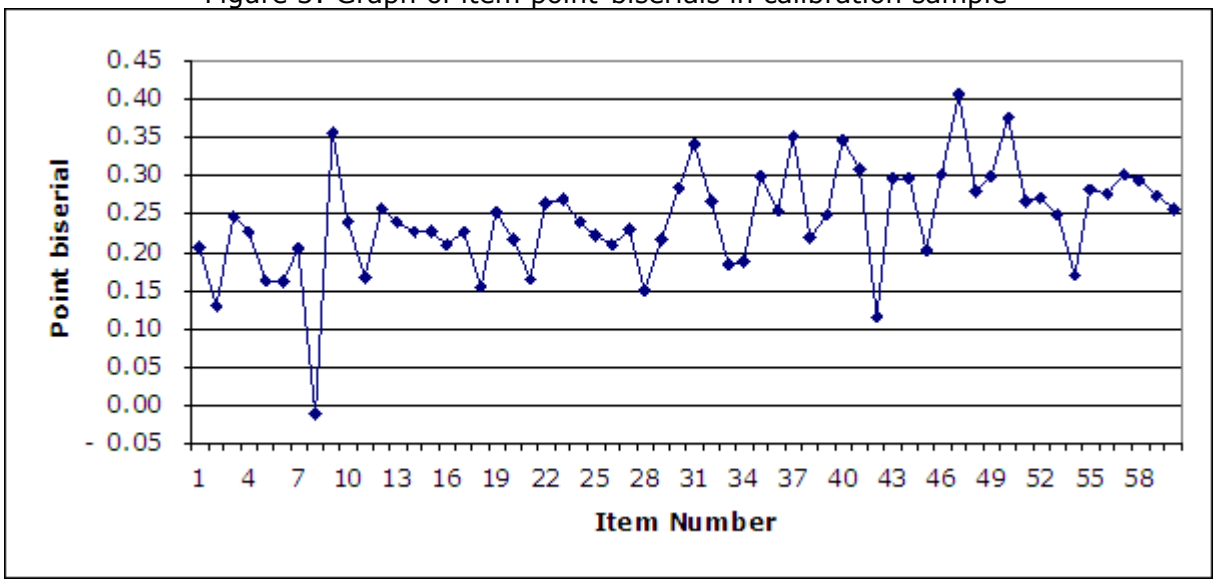
Partitioning the data set

The Illinois IRT Modeling Lab stresses the advantages of using cross-validation methods when fitting IRT models to data. In Rasch scaling we see if the data will fit the model, not vice versa, but nonetheless the matter of cross-validating findings is always recommended when sample size permits.

Before I started my work, I used a standard Lertap option called "To halve and hold" to randomly divide the data set into two equal parts. One of these became the calibration sample, while the other was held in reserve, later becoming the cross-validation sample. There were just under 1,600 cases (students) in each part.

The matter of item discrimination is focal to this study, and, accordingly, I have plotted the point-biserial coefficients in the calibration sample below, in Figure 3.

Figure 3: Graph of item point-biserials in calibration sample



Item 8's point-biserial value in the calibration sample was -0.01; the two highest point-biserial values were item 47 (0.40), and item 50 (0.38).

Other relevant data in the calibration sample: average score was 41.05; standard deviation 7.57; low score 18; high score 60 (100% correct); coefficient alpha 0.82; standard error of measurement 3.20; and the conditional standard error of measurement at 1.57 standard deviations above the mean was 2.30.

ConQuest 2.0

ConQuest is regarded as the successor to QUEST. It does much more than the simple Rasch scaling which is the focus of this paper. To get an idea of the scope of ConQuest, I suggest a look at its manual, available as a free download from www.assess.com. (There should be a dedicated website for a product of this nature, but as of April 2008 there was none. It was not even a straightforward matter to find an email address for the ConQuest development team, there being none in the manual. However, I was able to obtain an address by writing to www.assess.com, and once I had it I found the help support to be very good.)

ConQuest requires input to have an ASCII, or text file, format. Lertap will produce such files. It has an option to produce suitable data files for Bilog MG, and for XCALIBRE. Usually one of these file formats may be very readily adapted to meet the needs of other programs.

In this case I had Lertap output an XCALIBRE-ready "DAT" file for the calibration dataset. Lertap puts four header rows at the start of its XCALIBRE data file. One of them, the string of correct answers, can be copied to the "CQC" control file required in order to run ConQuest; after this all four lines should be deleted from Lertap's DAT file, leaving the file ready for immediate use by ConQuest. The following control lines were given to ConQuest 2.0:

```
Datafile EEQT93DS2bCal.DAT;
Format id 1-4 responses 5-64;
Key 2433221223441134214334111334411444222331143113334242224444113 ! 1;
Model item;
Estimate ! nodes=50, stderr=full;
Show >> EEQT93DS2bCalNodes50.shw;
```

The EEQT93DS2bCal.DAT file referenced above was the XCALIBRE DAT file produced by Lertap after the four lines mentioned were deleted.

ConQuest went through 350 iterations before stopping, finding "The maximum change in the estimates is less than the convergence criterion". The convergence criterion used was the default value of 0.0001. My Windows XP computer required about four minutes to get to this point.

The Show >> command produced a text file with item statistics. Results for the first ten items are seen in Table 3.

Table 3: Item statistics items 1 - 10

			UNWEIGHTED FIT			WEIGHTED FIT		
item		ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1	1	-0.340	0.062	1.01	(0.93, 1.07)	0.2	1.02	(0.94, 1.06)	0.5
2	2	-1.029	0.075	1.08	(0.93, 1.07)	2.1	1.03	(0.90, 1.10)	0.6
3	3	1.610	0.055	1.02	(0.93, 1.07)	0.7	1.01	(0.96, 1.04)	0.3
4	4	-0.660	0.067	0.98	(0.93, 1.07)	-0.4	1.00	(0.92, 1.08)	0.0
5	5	-1.651	0.095	0.95	(0.93, 1.07)	-1.3	1.00	(0.85, 1.15)	0.0
6	6	1.373	0.054	1.09	(0.93, 1.07)	2.4	1.07	(0.97, 1.03)	3.9
7	7	1.714	0.056	1.06	(0.93, 1.07)	1.7	1.03	(0.96, 1.04)	1.4
8	8	1.956	0.058	1.25	(0.93, 1.07)	6.6	1.15	(0.95, 1.05)	5.8
9	9	1.221	0.053	0.95	(0.93, 1.07)	-1.5	0.95	(0.97, 1.03)	-3.0
10	10	0.317	0.055	1.01	(0.93, 1.07)	0.3	1.02	(0.96, 1.04)	0.8

What to make of these results? How to decide if the data fit the Rasch model? This is a big question; the Rasch literature does not evidence convergence with regard to suitable fit criteria. Is the answer to be found in the results above?

The ConQuest 2.0 manual (Wu et. al, 2007, p.23) suggests that it is: "*For the MNSQ fit statistics we provide a ninety-five percent confidence interval for the expected value of the MNSQ (which under the null hypothesis is 1.0). If the MNSQ fit statistic lies outside that interval then we reject the null hypothesis that the data conforms to the model. If the MNSQ fit statistic lies outside the interval then the corresponding T statistics will have an absolute value that exceeds 2.0.*"

The ConQuest manual suggests that the |T|>2 criterion is best applied to the WEIGHTED FIT values in Table 3; these values are also known as "infit" figures by the Rasch community, while ConQuest's UNWEIGHTED FITs are "outfits". Were I to apply this guideline, thirteen of my sixty items would be said to misfit the model, including three of the items in Table 3: items 6, 8, and 9. (Others: items 16, 18, 21, 28, 31, 33, 37, 40, 47, and 50.)

The |T|>2 criterion in the ConQuest 2.0 manual matches that found in Chapter 3 of Bond and Fox (2001).

However, Chapter 12 of the same text, Bond and Fox (2001), has other guidelines; in Table 12.6 (p.179), we read that we can apply a range of 0.8 to 1.2 to both infit and outfit mean square values for high-stakes multiple-choice tests. Items whose MNSQ values are within this range may be said to fit the model. Were this criterion applied to my results, all sixty items would pass.

Then, still in Bond and Fox (2001), we read "values between 0.70 and 1.30 are generally regarded as acceptable" (p.230). My sixty items pass again, but then, just a few sentences later, we read that infit T values in the range -2 to +2 are "usually held as acceptable", which gets me back to having thirteen misfits since I have that number of items with WEIGHTED FIT values outside this range for T.

Above I mentioned that I also used Winsteps software; its help system suggests looking at outfit mean square values rather than infits, using a range of 0.5 to 1.5 as the criterion for judging fit. All sixty items pass; none of my items had UNWEIGHTED FITs falling out of this range.

Wu and Adams (2008) state that "*Many textbooks or other resources make recommendations on the range of acceptable mean-square values or t values for residual based fit statistics. There are probably no right or wrong answers. You will need to understand the issues with these fit statistics when you apply rules of thumb*." A bit later in the same document, Wu and Adams add that "... *when residual based fit statistics show that items fit the Rasch model, this is not sufficient to conclude that you have the best test. The reliability of the test and item discrimination indices should also be considered in making an overall assessment* ...."

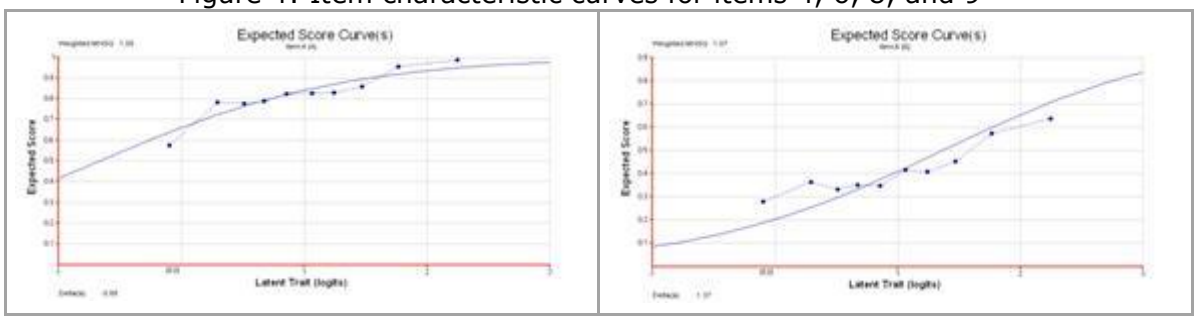
This matter of having an adequate fit and not being able to conclude that I have "the best test" was a comment which intrigued me; I sent off an email message to Wu. She replied that it is possible for totally random item responses to fit the Rasch model (Wood demonstrated this in 1978; more recently, Garcia-Perez (1999) provided related comments). One would expect random item responses to result in item discrimination indices centred on zero; coefficient alpha would also be expected to be zero. In other words, we could have nonsense classical item and test results and yet find the data to fit the Rasch model.

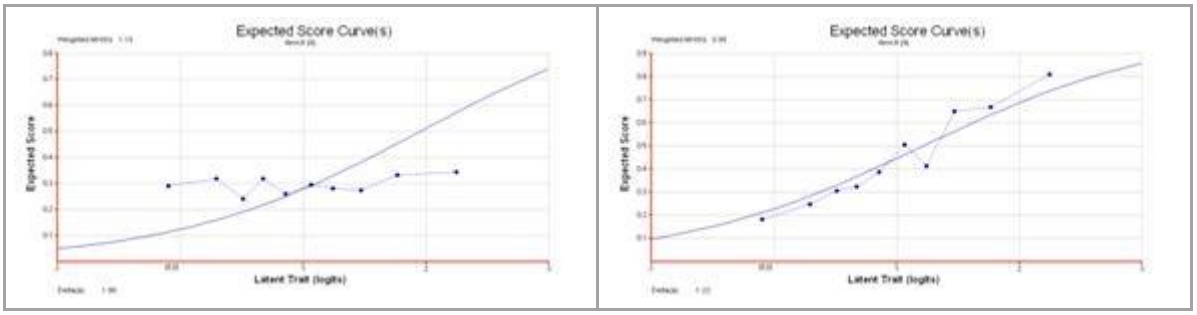
Another Wu and Adams (2008b) document suggests that a "good" item will have a high point-biserial discrimination ("*say above .4*"), a fit mean square close to one, and an observed item characteristic curve which is "close to the theoretical one". They also state that "*The reliability of an instrument is often used to judge the overall quality of the instrument*". We have always known this, of course, but Wu and Adams are using this statement in the context of Rasch scaling.

Like Winsteps, ConQuest 2.0 will calculate the discrimination and reliability indices found in classical test theory, such as those routinely computed by Lertap. However, the version of ConQuest I used did not correct its point-biserials for part-whole inflation; we might think, then, that their "say above .4" guideline for item discrimination might come down to about .3 or so.

Figure 4 shows four item characteristic curves, as created by ConQuest. The abscissa (the x-axis) is "Latent Trait (Logits)". Item 4 is plotted upper left; item 6 upper right; item 8 lower left; item 9 lower right. Infit MNSQ (WEIGHTED FITs in Table 3) values are given in the plots, as are item 'deltas' (ESTIMATEs in Table 3); the plots are well formatted, although obviously they are difficult to see here as I have reduced the size of each.

Figure 4: Item characteristic curves for items 4, 6, 8, and 9





Wu and Adams suggest we look at these curves as one means of seeing how well data fit the Rasch model. I admit to lacking practice in this regard, but it would not take much to conclude that item 8's fit is very poor (lower left curve in Figure 4). The only plot which seems suitable to me might be that for item 6 (upper left), but no doubt I should point again to my inexperience here.

As Rasch advocates will tell us, real data will seldom, if ever, perfectly fit the model. We should not expect to look at item response plots and see a perfect match between the sometimes-jagged lines connecting observed values and the always-smooth curves returned by the theory.

The critical question is: How much departure from a perfect fit should we allow? There are differing guidelines. In the case of my sixty-item exam, some guidelines suggest thirteen items do not fit. Others guidelines say all items seem to fit the model.

This is a major point. I am going about this business as I want to convert raw test scores to the Rasch scale with its purported interval measures. How will I know if this process is justified? If the data do not fit the model, ideally the software should tell me. Not only should it tell me, but it might go so far as to actually refrain from producing scale scores when it has found a poor fit.

My experience with ConQuest, and later with Winsteps, indicates that this does not happen. The software leaves it to me to decide. It will go ahead and make Rasch scale scores no matter how the data fit or do not fit the model, trusting me to be the final arbiter. There may be some room for concern here. The guidelines for judging fit are far from conclusive. They are loose enough to give test developers room to maneuver – some might reject items using one of the guidelines, while other developers, using different guidelines, could end up accepting all items. The decision rules lack focus. There ought to be something better, procedures or methods which are more objective.

I thought it might be useful to look at measurement precision, thinking that perhaps indicators of precision might serve as a workable fit criterion. Say I pick two scores, the average logit (log odds unit) score, and that logit corresponding to 1.57 standard deviations above the average. These are two points of interest along the score scale mentioned earlier, when details of the raw scores were discussed. What will standard errors of measurement look like if I try to fit all sixty items to Rasch, and then remove the thirteen?

With all sixty items, ConQuest computed the average logit to be 1.00, standard deviation 0.62. At the average logit score, the standard error was reported to be 0.31. At about 1.57 standard deviations above the mean, the standard error was 0.34. Converting the logits to a scale with mean 500, standard deviation 100: standard error at the mean would equal 49.86, while further along the scale, at about 657, the standard error would be 59.90.

Next I took out the thirteen items. ConQuest went through 237 iterations before converging. None of the forty seven items had infit T values with a magnitude at or over 2. The average logit score was 1.22, standard deviation 0.63. Standard error at the mean logit was 0.36; at about 1.57 standard deviations out the standard error was 0.42. Converting these logits to a scale with mean 500, standard deviation 100: standard error at the mean would equal 57.23, while out at a scale score of 657 the standard error would be 66.77.

Clearly there has been no gain in taking out the thirteen items if measurement error is a criterion. In fact, blind adherence to the |T|>2 guideline has not served well – a closer look reveals that the guideline has taken out six items with low point-biserials (6, 8, 18, 21, 28, 33), which seems acceptable, but it has also wiped out six of the most discriminating items (9, 31, 37, 40, 47, 50). Test reliability has decreased to 0.78; measurement error has increased.

The Rasch model postulates equal discrimination for all test items. The |T|>2 guideline appears to have jettisoned the test items whose point-biserials differ most from the average point-biserial value. More experienced Rasch users might have expected this.

Wu suggests that items with high point-biserials should not be eliminated; this is one of her selection criteria. (Note that such items will have negative MNSQ values, while items with the lowest point-biserials will have positive MNSQs, making them easy to identify. Negative MNSQs indicate items with better than expected fit. Linacre and Wright (1994, p.350) suggested retention of such items even when T<-2.)

My next step was to filter out the six items with low point-biserials (6, 8, 18, 21, 28, 33). Test reliability came back up to 0.82. With fifty-four items to work with, ConQuest went through 286 iterations before converging. Seven of the fifty-four items had infit T values with a magnitude at or over 2 (items 7, 9, 15, 16, 40, 47, 50). The average logit score was 1.15, standard deviation 0.67. Standard error at the mean logit was 0.33; at about 1.57 standard deviations out, the standard error was 0.43. Converting these logits to a scale with mean 500, standard deviation 100: standard error at the mean would equal 49.00; at a scale score of 657 the standard error would be 63.84.

This matter of moving items in and out of the picture demonstrated a couple of things. First, application of the |T|>2 guideline was not useful at all as it resulted in a decrease in measurement precision. Stepping back a bit and looking at things from a classical test theory perspective had me taking out six items whose point-biserials were among the lowest of the lot (positive T value). This appeared to restore measurement precision around the test average, but not so out at 1.57 standard deviations from the average.

At this stage I began to wonder about the possible effects of misbehaving students. The Rasch model is sensitive to students whose response patterns upset the model. Rasch neither requires nor assumes Guttman-type response patterns, but it is the case that the top students should do better on the hardest items, where the weak students are expected to struggle. Should there be excessive deviance from this general pattern, strong students missing easy items, weak students doing well on hard items, we might say that the responses are "noisy", and not blame the Rasch model for suggesting that the data have a poor fit. (In this context, items might be said to be noisy when their discrimination values are out of line with other test items.)

The April 2008 version of ConQuest did not provide indicators of student fit. I put ConQuest aside, compared the amount of funds in my budget to the prices of other Rasch systems, and purchased Winsteps[1].

[1] A basic working version of Winsteps software is included with Bond and Fox (2007).

Winsteps

I had used an XCALIBRE-formatted DAT file from Lertap with ConQuest. This format allows me to use the original response codes; ConQuest then requests a key for each item so that it can score the responses. When the user has entered original response codes, ConQuest's plots are able to graph the response patterns, providing a visual means for interpreting the performance of item distractors, much as Lertap's quintile plots do.

Winsteps will also accept original responses as input, with a companion key string of correct answers. Winstep provides data formatting aids, including one which supports importing data from an Excel worksheet, and another which does a fair job of adding control records to the top of an ASCII data file.

I tried these aids, but ended up copying and modifying a sample ASCII file which came with the program. This example assumed that the items had been pre-scored, so I had Lertap create a Bilog-formatted `DAT` file from the sixty-item calibration sample, and then added the following control lines to it.

```
&INST
TITLE = "EEQT93 DS2b calibration sample."
PERSON = Person ;
ITEM = Item ;
ITEM1 = 6 ;
NI = 60 ;
NAME1 = 1 ;
NAMELEN = 4 ;
XWIDE = 1 ;
CODES = 01 ;
UIMEAN = 0 ;
USCALE = 1 ;
UDECIM = 2 ;
MJMLE = 0
LCONV = .001
RCONV = 0
CONVERGE= Logit
KEY1 = 111111111111111111111111111111111111111111111111111111111111 ;
ptbiserial=y
&END
```

Winsteps was much faster than ConQuest, taking a matter of just seconds to converge. Once it had calibrated the model as best it could, given my sixty-item test data, Winsteps remained in a console-like mode, waiting for me to request tables and / or output files.

The first thing I wanted to do was compare Winsteps' output with that from ConQuest. A scan of item measures and their standard errors revealed close agreement between the results from the two systems.

I then applied the |T|>2 cutoff to the MNSQ values from Winsteps. Fifteen items failed this test: the thirteen identified by ConQuest, plus two more, items 46 and 49.

Were I to apply another of the guidelines used above, looking for mean square values beyond the 0.8 to 1.2 range, four items would be highlighted: item 8 with an outfit mean square of 1.35, item 42 with an outfit of 0.78, item 50 with an outfit of 0.79, and item 51 with an outfit of 0.70.

Winsteps has its own suggested guidelines: accept items if their outfit MNSQ values are within the range of 0.5 to 1.5. This would result in accepting all items, as may be seen in Figure 5. (Note that when a figure such as Figure 5 is produced by Winsteps, the

result is easy to read, and colourful. Here the figure has been resized and the colours removed in order to fit the requirements of this journal.)

Figure 5: Item bubbles

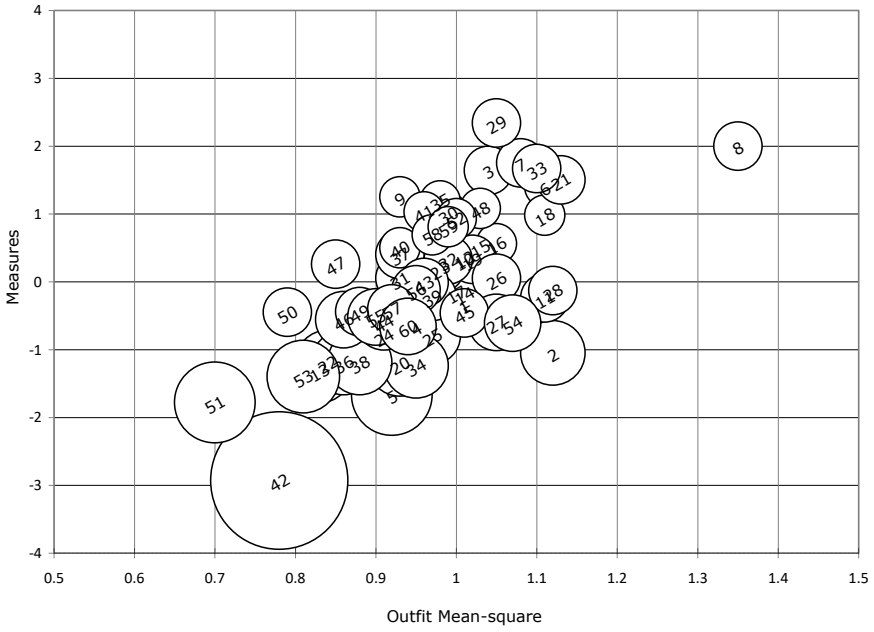


Figure 5 is an Excel "bubble chart". Winsteps passes item data to Excel, along with the prompts necessary to get Excel to make a bubble chart from the data.

The plot in Figure 5 resembles the "pathway" common in Rasch scaling. Logits are plotted along the y-axis, the "Measures" axis. Outfit mean squares comprise the x-axis; this axis is usually centred on a value of 1 (one), the expected mean square value for items when data fit the model. The centre of each circle displays an item number; the size of the circle reflects the relative[2] magnitude of the item's standard error. The most difficult items have their circles at the top of the chart. In this case item 29 was the most difficult; its Rasch measure was 2.34, with standard error .06. Item 42's bubble is lowest on the chart. This was the easiest item in the test, with a Rasch measure of -2.93, standard error of .17.

Item 8's circle is the right-most bubble on the chart, indicating that it had the largest outfit mean square value (1.35). It was also a difficult item, with a Rasch measure of 2.00, standard error .06.

Ordinarily the y-axis in this sort of chart rises from the usual centre of the mean squares, that is, from 1 (one) on the x-axis. When it does it is seen as representing the logits "path", the heart of a Rasch scale.

Figure 6 has the same sort of chart, this time for the 1,600 students in the calibration sample. With so many bubbles the chart is obviously difficult to read[3]; however, as with

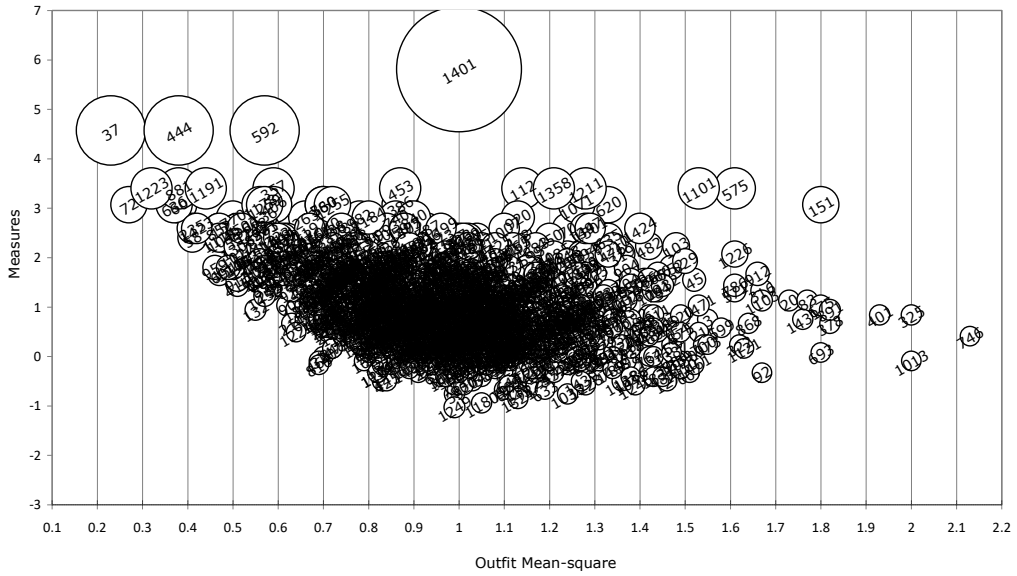
[2] The size of the bubbles may be readily changed in Excel. I have left bubble size at its default Winsteps setting for the charts seen in this paper. In general the default setting results in bubbles with standard errors seemingly much greater than they really are, but this makes it easier to visually compare the standard errors.

[3] A reviewer suggested that the bubble charts are sometimes easier to interpret when the bubble size is reduced.

the item bubbles in Figure 5, relative outliers may be seen – the top bubble in Figure 6 corresponds to student 1401, the only student to get a perfect score on the test.

A problem with Rasch scales is that measurement becomes imprecise at the extremes, making for larger bubbles, such as that seen for 1401. In fact, the model is not capable of estimating measurement error for perfect scores, of for imperfect ones; the radius of student 1401's circle is really infinity. This matter is further discussed below.

Figure 6: Student bubbles



Combining Figures 5 and 6 is all that is necessary to make a person and items "map" somewhat similar to the "Bond map" seen in Bond and Fox (2007, p.53). Winsteps will readily do this, but the result, when seen in black and white, is even more difficult to take in than is Figure 6. Winsteps will also output a text-based map of persons and items, similar to the Wright map pictured in Bond and Fox (2007, p.55)[4].

Person and item maps can be handy, especially when we think of a developmental pathway for cognitive growth. Both editions of Bond and Fox (2001, 2007) have particularly compelling examples of their use (http://homes.jcu.edu.au/~edtgb/book/); the manuals for some of the Progressive Achievement Tests also nicely exemplify the use of such maps (http://www.nzcer.org.nz/default.php?products_id=1553). The "steps" in Winsteps (and now in Bond&FoxSteps, the software which accompanies Bond and Fox (2007)) refer to progression along the path. Individual items appear at various spots on the path, and when student scores are superimposed we get an idea of where each student may be, how many "steps" he or she has taken, and which test items may have been stepped over in the process.

The Winsteps help system suggests that students whose outfit mean squares exceed 2 in magnitude may serve to "degrade measurement". Student 746, the right-most bubble in Figure 6, was one, having an outfit mean square of just over 2.1. Winsteps found several of this student's responses to be "very unexpected". Given an estimated measure of 0.41, this student got too many easy items wrong (including item 42), and one very hard item correct (item 8). The model allows for some variations of this sort, but Winsteps judged this response pattern to be aberrant.

[4] I attempted to obtain permission to use some of the figures in Bond and Fox, but gave up after seeing the publisher's extensive, detailed authorisation process. Passports are easier to get.

There were three others with outfit mean squares beyond or on the edge of the suggested limit. I eliminated these three, along with student 746, and got new results. Finding it an easy matter to remove students and items from Winsteps analyses, at this point I also went on to repeat some of the ConQuest-based work reported above.

Results are shown in Table 4.

The row in Table 4 with "60<4S" under the No. Items column corresponds to the Winsteps output regarding all sixty test items, with four students removed. The S.E. (mean) column has standard error figures noted at the average logit score. The S.E. (1.57) column has standard errors for measures centred at 1.57 standard deviations above the logit mean. The "Normed to 500, 100" columns contain corresponding standard error values on a scale with mean 500, standard deviation 100. Note that "normed" is a label used by Winsteps – I would instead call this a "standardized" scale as the test results have definitely not been normed in the classic sense, nor have they been normalized.

Table 4: Estimates of measurement precision

	Logits calculated by Winsteps				Normed to 500, 100	
No. Items	Mean	S.D.	S.E. (mean)	S.E. (1.57)	S.E. (mean)	S.E. (1.57)
60	1.03	.79	.31	.43	39	52
47	1.27	.88	.37	.52	42	60
54	1.19	.89	.34	.48	38	54
60<4S	1.03	.80	.31	.42	39	52
				Lertap (60 items)-->	44	30

Table 4 summarises results concisely: as seen earlier with ConQuest, removing the thirteen items suggested by the |T|>2 guideline results in less measurement precision (the standard errors are higher in the row with 47 items). However, the row with 54 items shows that removing six items with low point-biserials does not greatly affect precision. The 60<4S row indicates that removing four students with excessive outfit mean square does not have much impact either (perhaps to be expected with a sample size of about 1,600 students).

The Lertap (60 items) values seen in Table 4 were computed for the true-score scale using the compound binomial model; there is more about this below (see Figure 7).

Dimensionality revisited

Linacre (1998) wrote: "*An ideal of the Rasch model is that all the information in the data be explained by the latent measures. Then the unexplained part of the data, the residuals, is, by intention, random noise."*

Winsteps has routines for examining residuals; see Linacre (2003).

The results of Winstep's analysis of residuals in the calibration sample are shown below in Table 5. In this case the Rasch model's measures account for about 40% of the total variance. Ideally this might be higher, but the analysis also indicates that there seems to be no systematic source for the unexplained variance, that is, no meaningful factor or component, just noise. I would conclude that these results support the idea of a single factor or component, suggesting unidimensionality.

Table 5: Winsteps residuals analysis, calibration sample

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)					
			Empirical		Modeled
Total variance in observations	=	2595.9	100.0%		100.0%
Variance explained by measures	=	1025.9	39.5%		38.6%
Unexplained variance (total)	=	1570.0	60.5%	100.0%	61.4%
Unexplned variance in 1st contrast	=	46.6	1.8%	3.0%	
Unexplned variance in 2nd contrast	=	45.6	1.8%	2.9%	
Unexplned variance in 3rd contrast	=	40.9	1.6%	2.6%	
Unexplned variance in 4th contrast	=	39.8	1.5%	2.5%	
Unexplned variance in 5th contrast	=	38.0	1.5%	2.4%	

The table seen above is complemented by eigenvalue scree plots in the Winsteps output.

I wrote to the Winsteps help desk to ask why the empirical variance explained by measures (39.5%) was greater than the modelled equivalent (38.6%); an answer was returned quickly: "in this case ... there is a source of general dependency within the data which constrains the data to slightly overfit" (Linacre, personal email correspondence, 29 April 2008).

It initially seemed that Winstep's analysis of residuals was an effective method for looking at the question of dimensionality. I was set to recommend it. However a newer version of Winsteps, version 3.65.0, was released while I was working on this paper. I upgraded to this version, and once again looked at a principal components analysis of residuals. Results differed much more than expected; see Table 6.

Table 6: Winsteps 3.65.0 residuals analysis, calibration sample

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)					
			-- Empirical --		Modeled
Total raw variance in observations	=	2031.2	100.0%		100.0%
Raw variance explained by measures	=	461.2	22.7%		23.0%
Raw variance explained by persons	=	167.8	8.3%		8.4%
Raw Variance explained by items	=	293.3	14.4%		14.6%
Raw unexplained variance (total)	=	1570.0	77.3%	100.0%	77.0%
Unexplned variance in 1st contrast	=	46.6	2.3%	3.0%	
Unexplned variance in 2nd contrast	=	45.6	2.2%	2.9%	
Unexplned variance in 3rd contrast	=	40.9	2.0%	2.6%	
Unexplned variance in 4th contrast	=	39.8	2.0%	2.5%	
Unexplned variance in 5th contrast	=	38.0	1.9%	2.4%	

Compare Table 6 with Table 5. The variance explained by measures is further elaborated in Table 6 where it is broken into persons and items. This is a useful enhancement. However, the total variance in observations in Table 5 does not match what is called the total raw variance in observations in Table 6. The empirical variance explained by measures has come down from 39.5% to 22.7%; the unexplained variance has increased from 60.5% to 77.3%.

I wrote again to the Winsteps help desk, seeking explanation. Linacre (personal correspondence, 1 May 2008) quickly responded:

> *"The variance decomposition in 3.64.2 was based on my earlier, largely theoretical analysis. Experience has indicated that this was over-optimistic in the assignment of variance to the Rasch dimension. 3.65.0 corrects this, and also attributes variance to the person and item distributions".*

There is here an indication of unsettled software. Such a major change makes it very difficult to compare results from Winsteps 3.65.0 with those from studies which used earlier versions of the software. Comparing the variance explained by measures from

study to study might be a common inclination among researchers; a software revision of this magnitude makes this inadvisable. We will have to wait until more studies report their results based on the new methodology seen in Winsteps 3.65.0, and at the same time hope that the method stabilizes.

Measurement precision, and reliability

I have included estimates of measurement precision in this paper mostly in an effort to suggest an appropriate criterion for deciding how well data fit the Rasch model. It is clear that one needs a practiced eye, as Wu might say, when it comes to interpreting the mean squares, T values, and item characteristic curves which Rasch will produce.

When I applied the |T|>2 criterion to my sixty-item achievement test, thirteen items failed. However, removing these items from the test lowered reliability, and increased measurement error. Using a Wu-suggested procedure of taking out items with low point-biserial values saw me remove six items from the test without a real loss in reliability and measurement precision.

The relationships between item point-biserials, reliability, and measurement precision have been known for a long time. High point-biserials result in higher reliability and less measurement error.

These points have nothing to do with item response theory or Rasch scaling; they stem from CTT, classical test theory. Wu suggests that the application of CTT methods in Rasch scaling is appropriate, and likely to lead to better scales.

This makes sense. A Rasch scale is a monotonic transformation of raw test scores. The correlation of Rasch measures and raw test scores will always be close to one; similarly, the correlation between Rasch item location estimates and the item difficulty figures from CTT will also be close to one, although reversed (that is, close to minus one; in Rasch scaling items with higher location parameters are the more difficult items, but in CTT, items with higher difficulty figures are the easier items).

Thus an appropriate criterion for reflecting on the quality of a Rasch scale might very well be the reliability index from classical test theory. However, there is a problem: high classical test reliability does not mean that a test is unidimensional, and IRT methods, Rasch included, are known to be sensitive to dimensionality. Would-be Rasch users might be well advised to look at dimensionality; in this paper I have used eigenvalues to do so, reporting on those produced by two systems: Lertap and Winsteps. (Nelson 2005 makes other relevant points about eigenvalues, scree plots, and reliability.)

According to Adams (2005), measures of test reliability in item response theory have been "... *seen as less important than in traditional test theory*..." (p.164). But Adams goes on to note that "... *a majority of users require an index, such as reliability, as evidence of the quality of the test*...." (p.166). Rasch programs cater to this requirement in several ways.

The Winsteps program, like ConQuest, outputs the value of the reliability coefficient used by classical test theory. In Winsteps, this is referred to by a variety of terms, including Cronbach Alpha, KR-20, Person Raw Score Reliability, and Kid Raw Score Reliability. ConQuest 2.0 refers to this reliability as Cronbach alpha, and as KR-20.

Both programs also report "person separation reliability". Bond and Fox (2001) describe this as "... an estimate of how well one can differentiate persons on the measured variable ....The estimate is based on the same concept as Cronbach's alpha." (Also see Adams 2005.) There is a similar statistic for items, referred to as the "item separation reliability". According to the Winsteps help system, low values of item reliability

"indicate a narrow range of item measures, or a small sample". Both of these reliability indices, person and item, have a minimum possible value of zero, maximum of one.

Winsteps adds another two reliability categories, further distinction of the two separation indices: "real" and "model", with the latter also referred to as "Rasch reliability". These are said to be estimates of "true" reliability, that is, true person variance divided by observed person variance. "Real" reliability is the lower boundary to true reliability, with "model" (Rasch) reliability being an upper boundary – true reliability will lie between these values. Note that these terms are used in the context of Rasch scales, with measures being logits. Cronbach alpha is an estimate of reliability based on the true-score scale. Table 7 displays the values of the various reliability coefficients, as found by Winsteps. The values for Cronbach alpha agree with those obtained in Lertap.

Table 7: Reliability figures

		Person Separation		Item Separation	
No. Items	Cronbach alpha	Real	Model (Rasch)	Real	Model (Rasch)
60	.82	.82	.82	1.00	1.00
47	.78	.77	.78	1.00	1.00
54	.82	.81	.82	1.00	1.00
60<4S	.82	.82	.82	1.00	1.00

Raw scores and measures

IRT models, including Rasch, work over an unbounded latent trait scale, free to take on values between minus infinity and plus infinity. In Rasch scaling, logits are used along the scale. Rasch logits are often claimed to be interval measures, and are generally simply referred to as "measures". This is in contrast to raw test scores which are seen by some as being closer to ordinal measures. (Wu and Adams (2008) state: "*raw scores provide measures somewhere in-between ordinal and interval measurement*", and: "*raw scores, while not quite providing an interval scale, offer more than just ordinal scales*".)

Figure 7 graphs standard errors of measurement from Lord's Method IV (Lertap's 'CSEM2' index), and from Winsteps. The ordinate in the figure corresponds to standard errors divided by respective standard deviations, with the abscissa displayed over the true test score scale.

Figure 7: Comparison of relative measurement error

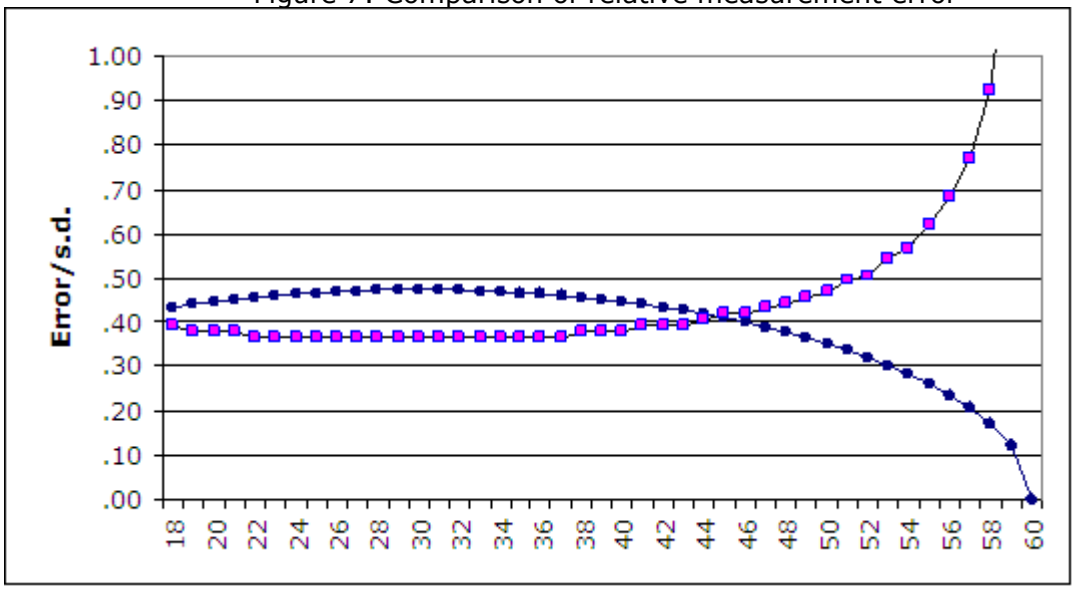


Figure 7 reflects what the Winsteps documentation refers to as "a paradox": extreme test scores have perfect precision, but extreme Rasch measures have perfect imprecision. That is, standard errors on the true score scale are concave down, going to zero at either end, while standard errors on IRT scales, Rasch included, are concave up, rocketing up at both ends in a marked manner. The right side of Figure 7 shows the Rasch standard error blossoming upwards rapidly, while the CSEM2 estimate from Lertap is dipping to zero.

This is not an anomaly of my sixty-item test. Such divergence has been recognized for many years (see, for example, Lord 1980, pp. 89-90). Brennan (1998) suggests that theta (latent trait score) "*can be viewed as a rather severe transformation*" of true scores "*that causes conditional SEMs in IRT to be larger at the extremes*". (For more on this apparent paradox, see Kolen, Hanson, and Brennan (1992), and Lee, Brennan, and Kolen (2000). Felt and Qualls (1996) state that empirical evidence indicates that measurement error does indeed decline rapidly at both ends of the true-score scale; Brennan (1998) also provides empirical support of this nature. Nelson (2007) has related data.)

In this case the mean of the raw test scores was 41, standard deviation 7.6. On the logit scale for these data, the mean measure was 1.0, standard deviation 0.8. The relative standard errors are fairly close in the proximity of the mean, with Winsteps showing better measurement precision (less error) in the range of three standard deviations below the mean to approximately half a standard deviation above the mean; after that Lertap's CSEM2 (Lord's Method IV) claims superior precision. In the region of a raw score of 53 (logit score 2.3), the suggested cut-score for the break between distinction and high distinction, the error/s.d. ratio for CSEM2 is about 0.30, compared to about 0.55 for Rasch (Winsteps). Higher scores have been measured with more precision on the true-score scale.

Recapping

Using software packages such as ConQuest and Winsteps to fit data to the Rasch model is basically a straightforward process, providing data have been formatted in the manner expected by the programs, and assuming the user is conversant with the use of text (ASCII) files.

However, looking for evidence of fit quality is not so straightforward. The guidelines are ambiguous. Wu has suggested the use of classical item and test statistics in conjunction with the guidelines, in part as an extension to the guidelines, and in part as a check against having poor-quality, random-like response data. We always know enough about our data to be aware of its quality, yes, but nonetheless it would be possible, for example, to mis-key the items, either by making innocent errors, or, perhaps, by incorrectly specifying where the item keys begin in a text file of input data. Such actions might well result in poorly-formed, random-like data. A common perusal of classical item statistics would likely reveal the problem, whereas a review of mean square fit values from a Rasch analysis might not.

Determining fit quality is but one aspect of the situation. Another is being aware of the consequences of poor fit. This is not clear in the literature I was able to find.

Adams (2005) pointed out that reliability, per se, historically did not have a central focus in Rasch scaling. He and others have written about reliability-like measures which are appropriate for Rasch scaling, and Winsteps and ConQuest incorporate them. I found the Winsteps output to be easy to use in this regard – the reliability estimates are a prominent feature of some of the Winsteps tables.

Both software packages give top billing to standard errors. Winsteps will rescale the logits readily, and carry standard error estimates to the new scale in the process.

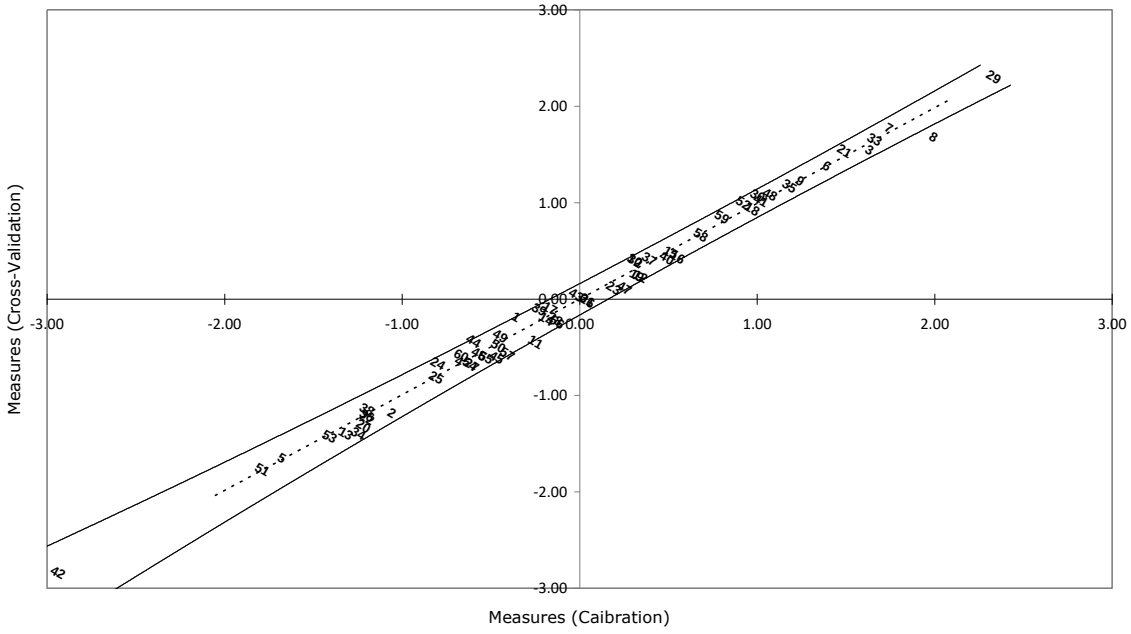
On comparing standard errors from Rasch to those derived by Lertap, using Lord's "Method IV", for the case of my sixty-item calibration sample, Rasch returned better relative precision in the centre of its measures. Lord's method, working on the true-score scale, resulted in more accurate score estimates away from the centre.

Cross validation

All of the results reported thus far have been based on a calibration sample randomly selected from the complete data set. I have held in reserve the cross-validation sample mentioned earlier, and will take a look at it now.

Winsteps facilitates the matter of comparing results from two data sets. Figure 8 is an example, plotting item measures in the two samples against each other, with 95% confidence interval bands. I realize that the figure is very difficult to read, but perhaps item 8's deviance can be detected – it is shown towards the upper-right of Figure 8, and is the only item to fall outside of the confidence bands. The item at the extreme upper right in this figure is number 29; that at the other end, an outlier in appearance, is item 42.

Figure 8: Calibration and cross-validation measures



The correlation between the item measures in the calibration set and the item measures in the cross-validation set was 0.996. Of interest is the correlation between item difficulties in the two sets: it was also 0.996. The correlation between item measures and item difficulties in the calibration sample was -0.979; in the cross-validation sample it was -0.981. (As a side comment, the correlation between item point-biserials in both samples was 0.864.)

I also used Winsteps to examine the residuals in the cross validation sample – see Table 8. Results were very similar to those reported earlier for the calibration sample, seen above in Table 5.

Table 8: Winsteps residuals analysis, cross validation sample

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)				
		Empirical		Modeled
Total variance in observations	= 2619.6	100.0%		100.0%
Variance explained by measures	= 1048.6	40.0%		39.3%
Unexplained variance (total)	= 1571.0	60.0%	100.0%	60.7%
Unexplned variance in 1st contrast	= 48.4	1.8%	3.1%	
Unexplned variance in 2nd contrast	= 44.6	1.7%	2.8%	
Unexplned variance in 3rd contrast	= 40.3	1.5%	2.6%	
Unexplned variance in 4th contrast	= 37.8	1.4%	2.4%	
Unexplned variance in 5th contrast	= 37.0	1.4%	2.4%	

Logits and interval scales

Very high correlations between item difficulty statistics from IRT and CTT are found repeatedly in the literature, and are in fact to be expected. I should also make mention of the correlation between student Rasch measures and their raw test scores: in the calibration sample it was 0.995; in the validation sample it was 0.983. This sort of result is also common, and again is to be expected. As Wu and Adams (2008) point out, "*Rasch scores do not alter the ranking of people by their raw scores. If one is only interested in ordering students in ability, or items in difficulty, then raw scores will serve just as well. No IRT is necessary*".

Why then, when we use achievement test data such as that found in this study, might we want to employ Rasch scaling? Well, there's the traditional advantage of IRT methods: measures of item difficulty and student proficiency are expressed in the same metric, on the same scale, or "pathway". The Bond maps and Wright maps sometimes seen in Rasch work are regarded by many as useful; the very readable books by Bond and Fox (2001, 2007) have good examples.

And then there's the matter of Rasch producing measures which are often considered to be on an interval scale. In this area the Rasch community derives particular pride, at times almost revealing an inclination to boast. In taking us from raw test scores (often but not always regarded as basically ordinal in nature) to linear Rasch measures, we are told to see ourselves as being more scientific, transported into a realm of objective, "fundamental" measurement.

Interval scales make it possible to read more into their values; we can interpret the numbers more exactly. For example: Jorge's Rasch logit measure on Qualifying Test EE130 is 2.42, Milagros' is 1.21, and Eduardo's is 0.00. Jorge is more proficient on the subject matter than Milagros, who in turn is more proficient than Eduardo. We cannot say that Eduardo has no proficiency at all; we do not have a ratio scale, we do not know what a score of zero means (indeed, to the left of Eduardo are others whose measures are negative). We cannot say that Jorge has done twice as well as Milagros; nor can we say that Jorge has rendered twice the proficiency. However, we do know, we can say, that on this proficiency scale, the distance in proficiency units between Milagros and Eduardo is the same as that between Jorge and Milagros.

A problem is that we do not know what these proficiency measures are. They represent a construct. Were we to use a common interval scale, time of day, with time A being 1200 hours, time B 1400 hours, and time C 1600 hours, we are not likely to have any problem of interpretation: the amount of time between C and A is twice that between A and B, and because we know this scale so well we use it readily. If I take my van for service at time A, I can watch three innings of baseball by time B, and about six innings if it will not be ready until time C (assuming no rain and a constant progression of innings; of course both B and C are imaginary times, the van is never ready until just before the garage closes).

Interpreting the time scale, an interval scale, is no problem. We grew up with this scale. The Rasch scale is much newer; we can be excused for finding its logits more difficult to interpret and use. But setting this aside, are we not better off with such measures?

Someone, perhaps the Dean herself, might well ask: In what sense, exactly? Our test scores may be plain, raw, possibly tending to ordinal, but here in our faculty we can interpret them with not much difficulty at all. We know our test very well. We trust it. We regard it as valid. We have examined the relationship between test scores and subsequent performance -- given the near-perfect correlations between Rasch measures and raw scores, this relationship would not be expected to change. We do not make Jorge-Milagros-Eduardo comparisons in your style; we are not sure what the proficiency scale means. If we tell our students that they got so many items correct, for a percentage score of such and such, they and their parents know what we mean.

Kline (2000, p.81) wrote: "*… despite the fact that items and individuals can be matched in Rasch scaling, norms are required to give meaning to the scores ….*". Then, regarding the use of Rasch scales over time, Kline (p.82) stated "*The claim that Rasch scales were* per se *meaningful and did not require norms was shown to be mistaken other than in a few specific cases and even here norms were likely to add meaning to the scales.*" (The specific cases Kline mentioned in his handbook had to do with developmental stages found in some mathematics and music curricula.)

In writing about a new approach to reporting NAEP[5] performance test results, Hattie (1999, p.414) wrote: "*The advantage of this definition of a market basket is the possibility of reporting students' performances in an observed-score metric that would avoid the need for the kinds of sophisticated statistical manipulations of data required by the current IRT scaling.*"

Rasch users will point out that rescaling logits is very common. The Winsteps program, for one, makes it easy to do. We do not have to have Jorge stuck fast to a Rasch measure of 2.42; we can rescale to anything we want, even to a scale which resembles that of the observed scores. We can do this without destroying the linear nature of the measures, if indeed that is their nature. (Bond and Fox (2007, p.206) used the term "user-friendly" in reference to this sort of rescaling, describing it as a means of converting logits to "more meaningful units".)

This may please some people, but there could be others who, like the Dean, are not convinced that we need to shift to Rasch scaling. Why go through this process only to end up with a scale which may well be more difficult for us to interpret than the one we use now? Because it is thought to be a linear scale? This sounds desirable, but its benefits seem muted if we do not make Jorge-Milagros-Eduardo comparisons. Now, if someone could come forth with a method which would produce a ratio scale, that would undoubtedly warrant the title of "fundamental measurement", and be likely to sway the Dean immediately (one would hope so).

Finally, and most importantly, the Rasch scale we may have laboured to produce in order to achieve an interval-level proficiency scale may in fact not have accomplished that at all. Citing work by Nickerson and McClelland (1984), Karabatsos (2001, p.393) wrote "*...it is possible for ... the Rasch model ... to conclude excellent or perfect data fit, even for data sets containing serious violations of the conjoint measurement axioms....*". In the same paper, Karabatsos (2001, p.395) wrote "*The Rasch model estimates give the illusion that the model can automatically construct additive conjoint measurement from any data set, no matter how noisy the data are. However, there is absolutely no basis to assume that such an automatic construction is possible*." (A readable account of additive

[5] National Assessment of Educational Progress (U.S.).

conjoint measurement is seen in Perline et al. (1979). Interval scaling depends on meeting the conjoint measurement axioms.)

Karabatsos's work, and that of Nickerson and McClelland, indicate that the extent to which Rasch scores are interval measures is open to question. Karabatsos (2001, p.420) wrote that "… *by definition, the IRFs under the Rasch model can only be data dependent, which leads to over optimistic conclusions with regard to the scalability of observed data sets* …." Karabatsos has developed software for use in testing the conjoint measurement axioms; his paper (2001) exemplifies its application to a few data sets. In personal correspondence (30 April 2008), Trevor Bond stated that "*A few are now trying to fathom the importance/impact of K's software*". Well they might; Karabatsos' work could be said to have revealed a fundamental flaw in a method thought to give us "fundamental measures".

Bond and Fox (2007, pp.273-274) suggested that data fit to the Rasch model "… *could now be considered fruitfully in terms of the extent to which actual response probabilities in any Rasch-modeled data sets violate the conjoint measurement axioms*…." They also wrote "*...it is the so far unresolved challenge ... to demonstrate that the procedures for determining whether the matrix of actual response frequencies adheres sufficiently to the Rasch measurement prescriptions really do satisfy these key conjoint measurement axioms. Rasch measurement is not there yet....*" (Bond and Fox, 2007, p.266).

There is room then to suggest more than one problem with Rasch scaling. The current guidelines for determining fit are sloppy. Measurement precision suffers markedly at scale extremes. Random responses have been shown to fit the Rasch model. Data which violate the basic assumptions of conjoint measurement have been shown to fit the Rasch model. An apparent good data fit to the model cannot be assumed to have resulted in interval measures.

At the beginning of this paper I referenced a quote from a conference paper: "Test scores are mere enumeration, a tally of number of items correct. They are not measures of proficiency. Rasch scales are objective fundamental measures expressed on an interval scale".

I would think that most Rasch experts would acknowledge that test scores are more than "mere enumeration". Well-developed achievement tests can almost always make a valid claim to be assessing subject mastery; on such tests there will indeed be a link between number of items correct and proficiency. That test scores may not be on a true interval scale is widely acknowledged. It seems, however, that there is scope to say that the same limitation might apply to Rasch measures, and this has certainly not been widely acknowledged at all.

References

Adams, R.J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31,* 162-172. (www.elsevier.com/stueduc)

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.)*.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Brennan, R.L. (1980). Applications of generalizability theory. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: The Johns Hopkins University Press.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22(4),* 307-331.

Crocker, L.M., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

du Toit, M.E. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Feldt, L.S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44,* 883-891.

Feldt, L.S., & Qualls, A.L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33(2),* 141-156.

García-Pérez, M.A. (1999). Fitting logistic IRT models: small wonder. *Spanish Journal of Psychology, 2(1),* 74-94.

Hanson, B.A., & Brennan, R.L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Hattie, J., Jaeger, R.M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 24,* 393-446.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2(4),* 389-423.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Kolen, M.J., Hanson, B.A., & Brennan, R.L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29(4),* 285-307.

Lee, W., Brennan, R.L., & Kolen, M.J. (2000). Estimators of conditional scale-score standard errors of measurement: a simulation study. *Journal of Educational Measurement, 37(1),* 1-20.

Linacre, J.M. (1998). Structure in Rasch residuals: why principal components analysis? *Rasch Measurement Transactions, 12:2,* 636.

Linacre, J.M. (2003). Data variance: explained, modelled and empirical. *Rasch Measurement Transactions, 17:3,* 942-943.

Linacre, J.M., & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8,* 350.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21(3)*, 239-243.

Nelson, L.R. (2000). *Item analysis for tests and surveys using Lertap 5.* Perth, Western Australia: Faculty of Education, Language Studies, and Social Work, Curtin University of Technology. (November 2006: www.lertap.curtin.edu.au)

Nelson, L.R. (2005). Some observations on the scree test, and on coefficient alpha. *Thai Journal of Educational Research and Measurement (ISSN 1685-6740), 3(1),* 1-17. (November 2006: copy seen at www.lertap.curtin.edu.au/Documentation/Techdocs.htm)

Nelson, L.R. (2007). Some issues related to the use of cut scores. *Thai Journal of Educational Research and Measurement (ISSN 1685-6740), 5(1),* 1-16. (November 2006: copy seen at www.lertap.curtin.edu.au/Documentation/Techdocs.htm)

Nickerson, C.A., & McClelland, G. H. (1984). Scaling distortion in numerical conjoint measurement. *Applied Psychological Measurement, 8,* 183-198.

Perline, R., Wright, B.D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3,* 237-255.

Qualls-Payne, A. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29,* 213-225.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103,* 667-680.

Wood, R. (1978). Fitting the Rasch model: a heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27-32.

Wu, M. (2008). *Model Fit*. Retrieved 27 March, 2008, from http://www.edmeasurement.com.au/Topic%20Four.doc

Wu, M., & Adams, R. (2008). *Applying the Rasch Model to Psycho-Social Measurement, a Practical Approach.* Retrieved 27 March, 2008, from www.edmeasurement.com.au/RaschMeasurement.pdf

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). *ConQuest 2: Generalized item response modeling software* [computer software]. Camberwell, Victoria: Australian Council for Educational Research. (March 2008: manual available from www.assess.com.)