Assessing the Invariance of Cognitive Item Statistics, an Empirical Study and Example Using Lertap5

Larry R Nelson[1]
Curtin University, Western Australia
Document date: 3 March 2020
website: www.lertap5.com
DOI: 10.13140/RG.2.2.36829.97769

Note

This document came about after Lertap5 had been updated with a new special macro, "RaschAnalysis1".

One recommended procedure for exploring how well item response data fit the Rasch model is to take a calibration sample, draw two random samples from it, and then compare results from the halves, looking for consistency, searching for invariance. It would be a good chance to stretch the new macro's legs.

The matter of Rasch parameter invariance was extensively demonstrated and advanced in a text by Wright and Stone (1979)[2]. The application of Rasch methodology was to provide an escape from the widely perceived sample-dependent limitations of CTT, classical test theory.

Was Fan's now oft-cited 1998 invariance paper one of the first chances for CTT to hit back with an invariance bat of its own[3]? It may have been. Nowadays it's easy to find a bevy of papers addressing invariance. Fan's findings have been supported; when the invariance wind blows, it can fan[4] CTT sails as much as it does IRT's.

Despite studies such as Fan's spilling at least some of the "uniquely sample-free" wind from IRT, it is still not uncommon to find IRT proponents appearing to feel that invariance still blows most strongly in their favour. I note, for example, that the creators of an IRT stalwart, Xcalibre, continue to promote IRT as being less sample-specific than CTT (as of February 2020). Might this be a subtle matter of sales over sails?

Below I compare the apparent invariance of cognitive item difficulty estimates using both CTT and Rasch IRT, finding essentially no difference with the data I chose to analyse.

We know now that findings like this are not novel. My intention in this paper is not to provide evidence of fresh research; rather, my objective is to suggest instructional content and examples that may perhaps be useful in classes having to do with test and measurement theory.

The vast majority of work seen in this paper is Excel based, with Lertap5 the main app. At times Lertap5 passes the baton to R and RMD scripts. There's an R and Lertap5 paper here.

---

[1] Comments / questions may be sent to l.nelson@curtin.edu.au
[2] Still a top read in my opinion.
[3] Fan found CTT statistics to be as invariant as IRT's.
[4] Pun intended.

Join the TAMILY?

Some, an ever-decreasing number, will know that we have not always had the internet. Our children and, especially, our grandchildren, may find that hard to believe, but years ago we shared datasets and even computer programs by exchanging magnetic tapes, hoping international post would deliver the goods without first submitting them to a metal scan (which might erase the tape).

Now, freely available datasets are often a matter of mouse clicks and we are fortunate indeed. Here's an outstanding example of a site with sample data covering a variety of topics possibly of use with classes.

In this document I'll use sample cognitive test items from TAM.

TAM Tutorial 3

I begin by looking at results gathered from the development and application of a 15-item mathematics test administered to 876 Year 7 Australian students. This paper provides a path to the actual dataset and corresponding test items. The discussion to follow begins with what I would use with classes and is often at quite a "fundamental" level. After presenting the material, and depending on questions they may have asked, I would likely see if students have followed along by asking them to repeat what I've done. This tends to work with class periods of two or three hours; one hour is a bit short.

CTT results

Classical cognitive test statistics are presented in Table 1, with corresponding scatterplot in Figure 1[5].

From a CTT perspective, this 15-item numeracy test has quite acceptable results. Item discrimination values are strong. Item difficulties are acceptable with perhaps a very minor observation being that there were no difficult items. For a 15-item test, coefficient alpha is very solid at 0.851.

I used a standard option in Lertap5 to create two random samples from the original dataset in order to look at the matter of "invariance" – will the item difficulty and discrimination statistics be similar in each of the random samples?

[5] Captured from a standard Lertap5 Stats1b report. Lertap5 may be downloaded here.

Options->	0	1	2	3	4	Difficulty	Discrimination
Q1	52%	48%				0.48	0.48
Q2	55%	45%				0.45	0.57
Q3		11%	6%	2%	81%	0.81	0.34
Q4		47%	34%	9%	11%	0.47	0.41
Q5	41%	59%				0.59	0.57
Q6		40%	29%	21%	10%	0.40	0.57
Q7	58%	42%				0.42	0.54
Q8	55%	45%				0.45	0.63
Q9	50%	50%				0.50	0.64
Q10	39%	61%				0.61	0.43
Q11		7%	6%	84%	4%	0.84	0.30
Q12		46%	22%	19%	13%	0.46	0.43
Q13	53%	47%				0.47	0.48
Q14	32%	68%				0.68	0.50
Q15	37%	63%				0.63	0.33

Table 1 (whole sample)

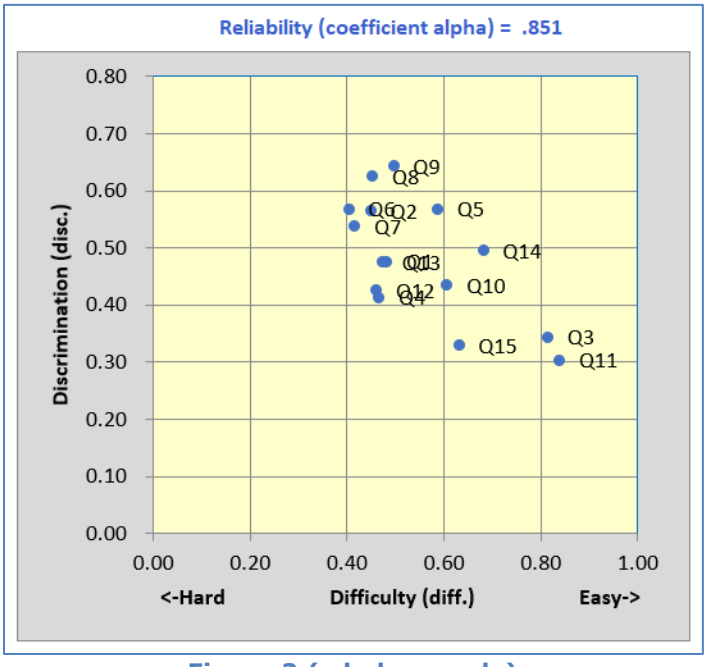

Figure 2 (whole sample)

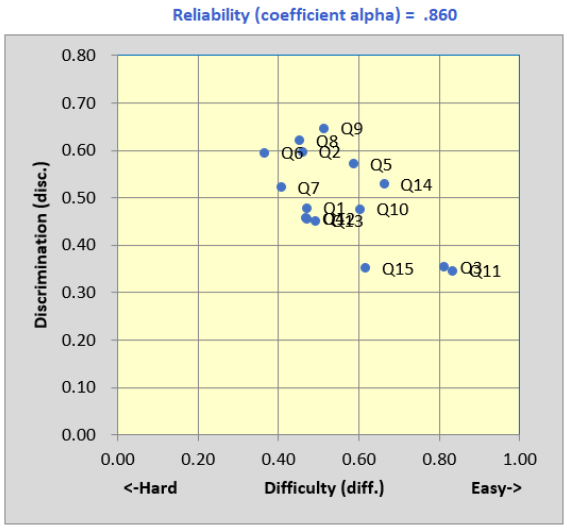

Figure 2 (random half1)

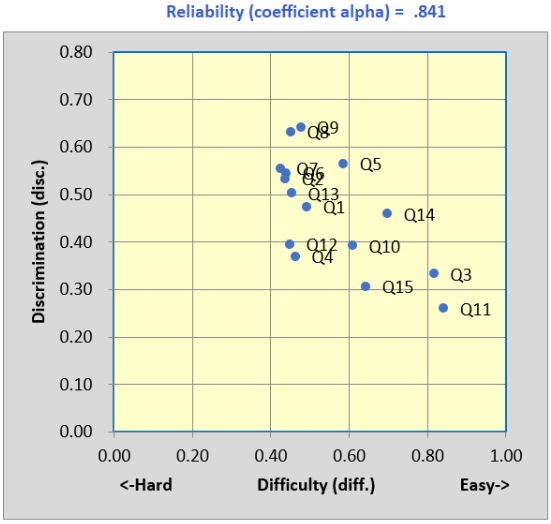

Figure 3 (random half2)

Figures 2 and 3 plot the distribution of item statistics using Sample 1 and Sample 2.

Are the statistics from the samples *identical*? No, clearly they're not. Table 2 makes it easier to compare results.

Q1 Half1	0.42	Q6 Half1	0.30	Q11 Half1	0.83
Q1 Half2	0.46	Q6 Half2	0.41	Q11 Half2	0.84
Q2 Half1	0.41	Q7 Half1	0.35	Q12 Half1	0.42
Q2 Half2	0.41	Q7 Half2	0.39	Q12 Half2	0.42
Q3 Half1	0.80	Q8 Half1	0.40	Q13 Half1	0.44
Q3 Half2	0.81	Q8 Half2	0.42	Q13 Half2	0.42
Q4 Half1	0.42	Q9 Half1	0.47	Q14 Half1	0.63
Q4 Half2	0.43	Q9 Half2	0.49	Q14 Half2	0.69
Q5 Half1	0.55	Q10 Half1	0.57	Q15 Half1	0.58
Q5 Half2	0.57	Q10 Half2	0.59	Q15 Half2	0.63

Table 2 (CTT Diff comparisons)

The item with the largest difference was Q6.

Figure 4 employs another way to look at the degree to which the difficulty statistics are similar. The plot in this figure was readily made using standard Excel chart options.

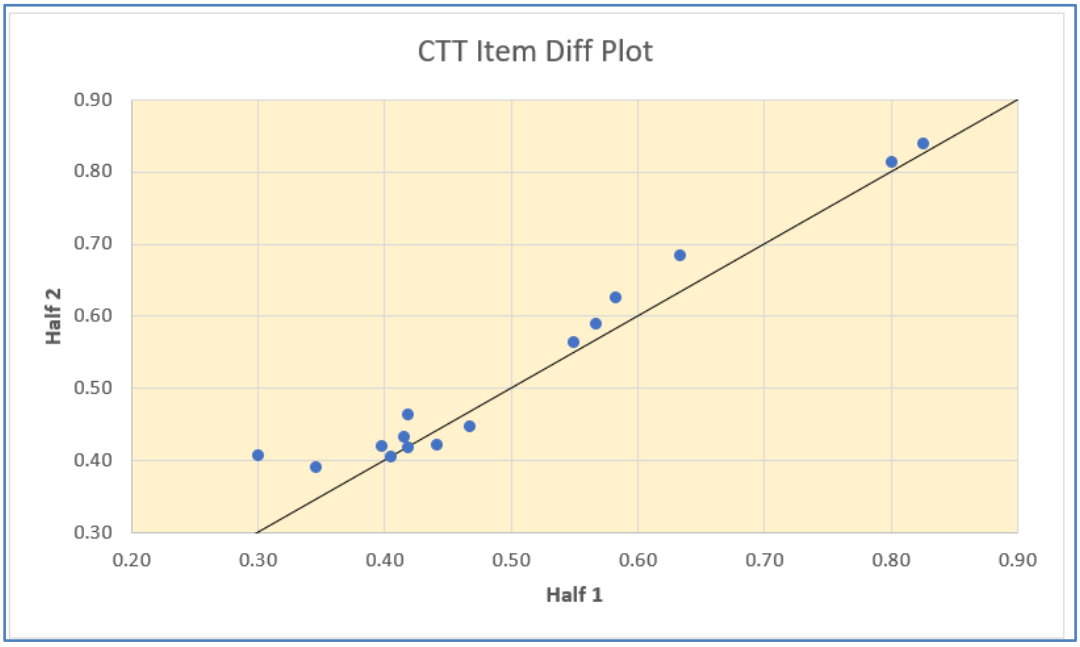

Figure 4

Each of the "blips" in Figure 4 represents one of the 15 test items. If the CTT Item Diff values were the same in both halves, each blip would be on the straight line. Thus, the degree to which the blips are distant from the line is an indication of variance – if they were all on the line itself there would be no variance; we'd then have "invariance".

The blip at the bottom left of the plot corresponds to item Q6 – the difficulty of this item was 0.30 in Half1 and 0.41 in Half2. Q14 was another item with a "blip" falling somewhat distant from the straight line – its CTT Diff value was 0.63 in Half 1, 0.69 in Half 2.

Fan (1998) used correlation coefficients as an index of the invariance of item statistics such as CTT Item Diff. The correlation was 0.978 for the Half1/Half2 figures plotted in Figure 4, a value quite similar to those seen in Fan's paper.

There is indeed great similarity between the two measures of item difficulty. They're not identical, no, but we could certainly say that they're very similar.

Are they "*invariant*"? No. They do vary. Not by much, but no, they are not exactly invariant.

IRT results

I followed the CTT analysis with a Rasch IRT analysis using Lertap5's "RaschAnalysis1" special macro. Results are displayed in Table 3.

We might expect the CTT Diff values in Table 3 to equal the Difficulty values seen in Table 1. They don't because students with test scores of zero and 15 have been excluded – in Rasch modelling, totally imperfect and totally perfect scores cannot be accommodated. Accordingly, responses for the nine students with zero test scores, and responses for the 70 students with perfect test scores, were not included when the macro ran. Results in Table 3 are based on 797 of the original 876 students.

Was application of the Rasch model appropriate? Do the data fit the model? A suggested way of answering these questions is to follow what I've done above with CTT: see if results from two random samples tend to evidence invariance. Should this turn out to be

the case, use of the Rasch results would appear to have justification. (Also see examples in Hambleton & Swaminathan (1991, Fig. 2.6), and Nelson (2008, p.19)[6].)

Item	CTT Diff	Rasch Diff	Error	Infit	Outfit
Q1	0.44	0.45	0.08	1.06	1.03
Q2	0.41	0.66	0.09	0.92	0.85
Q3	0.81	-1.80	0.10	1.02	**1.24**
Q4	0.42	0.55	0.08	1.14	**1.40**
Q5	0.56	-0.20	0.08	0.87	0.78
Q6	0.36	0.96	0.09	0.91	0.88
Q7	0.37	0.87	0.09	0.96	0.93
Q8	0.41	0.64	0.09	0.82	0.73
Q9	0.46	0.37	0.08	0.78	**0.67**
Q10	0.58	-0.32	0.08	1.07	1.08
Q11	0.83	-2.02	0.11	1.06	**1.41**
Q12	0.42	0.59	0.08	1.13	1.18
Q13	0.43	0.51	0.08	1.06	1.04
Q14	0.66	-0.79	0.09	0.92	0.87
Q15	0.60	-0.47	0.08	**1.23**	**1.40**
Average:	**0.52**	**0.00**	**0.09**	**1.00**	**1.03**
Median:	**0.44**	**0.45**	**0.08**	**1.02**	**1.03**
s.d.:	**0.15**	**0.90**	**0.01**	**0.12**	**0.24**

Table 3 (whole sample)

I applied the RaschAnalysis1 macro to the two random samples used for the CTT results discussed above. Results are seen in Table 4, in the Rasch Diffs summary in Table 5, and in the Rasch Diff Item Plot in Figure 5. In this case, the correlation between the Rasch Item Diffs was 0.979.

Half1				
Item	Rasch Diff	Error	Infit	Outfit
Q1	0.51	0.12	1.09	1.04
Q2	0.58	0.12	0.88	0.79
Q3	-1.83	0.14	1.05	1.16
Q4	0.53	0.12	1.11	**1.35**
Q5	-0.22	0.12	0.87	0.84
Q6	1.22	0.13	0.89	0.80
Q7	0.93	0.12	1.02	1.01
Q8	0.63	0.12	0.85	0.73
Q9	0.24	0.12	0.78	**0.66**
Q10	-0.32	0.12	1.02	1.07
Q11	-2.05	0.15	1.03	1.28
Q12	0.51	0.12	1.11	1.16
Q13	0.38	0.12	1.12	1.15
Q14	-0.71	0.12	0.91	0.83
Q15	-0.41	0.12	1.24	**1.39**
Average:	**0.00**	**0.13**	**1.00**	**1.02**
Median:	**0.38**	**0.12**	**1.02**	**1.04**
s.d.:	**0.91**	**0.01**	**0.12**	**0.22**

Half2				
Item	Rasch Diff	Error	Infit	Outfit
Q1	0.39	0.12	1.03	1.02
Q2	0.73	0.12	0.95	0.90
Q3	-1.77	0.14	0.98	**1.32**
Q4	0.57	0.12	1.18	**1.45**
Q5	-0.17	0.12	0.86	0.74
Q6	0.72	0.12	0.92	0.93
Q7	0.82	0.12	0.90	0.86
Q8	0.65	0.12	0.79	0.73
Q9	0.49	0.12	0.77	**0.67**
Q10	-0.31	0.12	1.11	1.10
Q11	-2.00	0.15	1.09	**1.52**
Q12	0.66	0.12	1.16	1.20
Q13	0.63	0.12	1.00	0.92
Q14	-0.88	0.12	0.93	0.92
Q15	-0.52	0.12	1.23	**1.43**
Average:	**0.00**	**0.12**	**0.99**	**1.05**
Median:	**0.49**	**0.12**	**0.98**	**0.93**
s.d.:	**0.89**	**0.01**	**0.13**	**0.27**

Table 4 (with the sample halves)

[6] References are here.

Q1 Half1	0.51	Q6 Half1	1.22	Q11 Half1	-2.05
Q1 Half2	0.39	Q6 Half2	0.72	Q11 Half2	-2.00
Q2 Half1	0.58	Q7 Half1	0.93	Q12 Half1	0.51
Q2 Half2	0.73	Q7 Half2	0.82	Q12 Half2	0.66
Q3 Half1	-1.83	Q8 Half1	0.63	Q13 Half1	0.38
Q3 Half2	-1.77	Q8 Half2	0.65	Q13 Half2	0.63
Q4 Half1	0.53	Q9 Half1	0.24	Q14 Half1	-0.71
Q4 Half2	0.57	Q9 Half2	0.49	Q14 Half2	-0.88
Q5 Half1	-0.22	Q10 Half1	-0.32	Q15 Half1	-0.41
Q5 Half2	-0.17	Q10 Half2	-0.31	Q15 Half2	-0.52

Table 5 (Rasch Diffs)

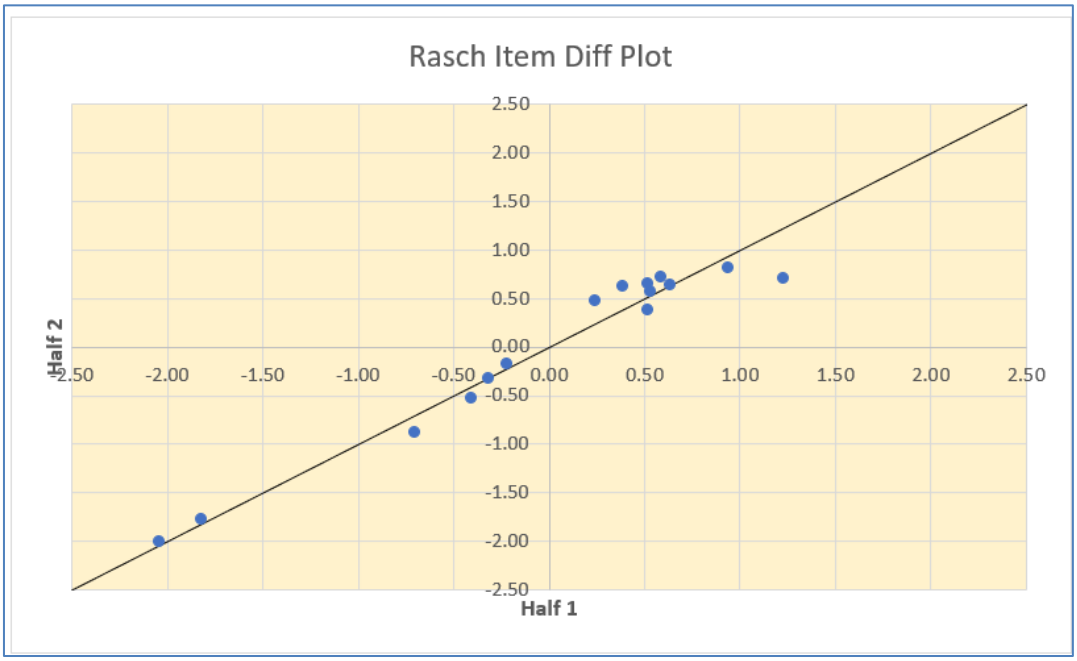

Figure 5

As was the case with the CTT results above, Q6 exhibited the greatest difference, having a Rasch difficulty of 1.22 in Half 1, and 0.72 in Half 2. In Figure 5, Q6 is the right-most blip, seen in the upper right quadrant, relatively distant from the straight line.

With regard to the question of "invariance", results computed using the two random samples have indicated that, be it CTT or IRT, item difficulty values vary from sample to sample – they're not invariant, but *they don't differ by a great deal in either case*. We could perhaps say that they seem to *tend* to invariance.

The results are basically the same for CTT and IRT in this example. Fan's 1998 study is supported here.

Discussion (possible pointers to share with students)

I might begin by seeing if I could startle students by saying that Figures 4 and 5 are "identical".

My intention would be to lead into a discussion of the differences in CTT/Rasch measure-

ment scales. I might begin by having them pick out the location of the Q6 blip in each of the figures. They should say "*The figures are not identical at all, what are you talking about (Sir), the Q6 blip is totally flipped around!"*

Now enter a possible surprize: if students have replicated my work, have them get into their two workbooks, Half1 and Half2 (or whatever they called them) and insert a new column to the immediate right of their original CTT Diff columns. The column will contain one minus the original CTT Diff value. For me, for example, Q6's original Diffs will become (1.00-0.30), and (1.00-0.41), or {0.70, 0.59}.

Then re-create Figure 4 using the two new 1-CTTdiff columns. Compare it to Figure 5. Are the figures now much closer to being identical (yes indeed!)? What's gone on here[7]?

Possible class exercises

I have used the "ToHalveAndHold" Lertap5 option to create two random samples, an option with a limitation: it always creates two samples of equal or near-equal size, and they may be larger than needed. If I want a random sample of, say, just 25% of the data records, I use an Excel-based program called "AbleBits" – not free but does have a 30-day trial period.

Even inexperienced Excel users should have little trouble using Lertap5, but may benefit from help when it comes to making the scatterplots seen in Figures 4 and 5 above. The "trick" is to get relevant Lertap5 output from the two samples to display side by side, as seen, for example, in Table 4.

In Table 4, data for each half were initially in separate workbooks called "Half1.xlsx" and "Half2.xlsx". These had been made using the ToHalveAndHold option. Within each of these workbooks, the "RaschItems1" worksheet, obtained by use of the RaschAnalysis1 special macro, had the needed results.

Table 4 may be assembled by copying the results from one workbook's RaschItems1 worksheet to the same worksheet in the other workbook. Figure 5 can then be made by selecting down the two "Rasch Diff" columns: starting with the "Rasch Diff" heading down to Q15's Rasch Diff value (-0.41) in one of the columns, holding down the <Ctrl> key on the keyboard and then selecting down the other Rasch Diff column. Do we not now have a plot that looks a lot like Figure 5?

With the two columns selected, a click on the "Insert" option from the Excel ribbon, followed by the "Recommended Charts" option, will get a "Scatter Chart" started.

What I've done to this point involves estimates of item difficulty. Item discrimination would be a possible additional topic, looking at the invariance of CTT values.

We could also compare point-biserial discrimination values with their biserial equivalents. Another IRT model might be employed, 2PL or 3PL, and its "a" parameter estimates compared with point biserials and biserials. These things can all be done in Excel with Lertap5 and with: (1) the experimental features option activated to get biserials, and (2) the handy, easy-to-use EIRT Excel add-in installed to get 2PL and 3PL output[8].

[7] Maybe this is too much work for what amounts to quite a simple point.
[8] EIRT only works with Windows computers. Mac users could get "a" estimates from the Itm package.

Dimensionality, DIF?

More challenging and enlightening exercises might look at dimensionality and differential item functioning.

Concerning dimensionality, if Fan's 1998 paper has been a recommended read, students will have seen the use of eigenvalues as an index of dimensionality. Lertap5 routinely outputs eigenvalues; tetrachoric correlations may be used in addition to the standard product-moment values, and scree plots are a cinch (Fig. 6).

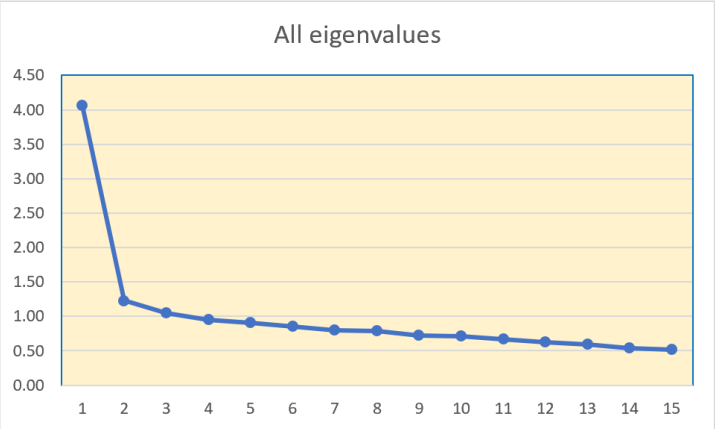

Figure 6

One can step up the action in this area by making use of Lertap5's Omega1 special macro to obtain a Schmid-Leiman solution (Fig. 7), accomplished by applying the **Psych** package in R. (This would obviously open a potentially vast area and may be appropriate only when students have had some prior exposure to factor analysis.)

```
Omega results from the R Psych package
Call: omega(m = MCQitems, digits = 3, title = '
    echo = FALSE)
Alpha:                 0.8
G.6:                   0.8
Omega Hierarchical:    0.57
Omega H asymptotic:    0.69
Omega Total            0.81

Schmid Leiman Factor loadings greater than  0.2
       g   F1*  F2*   F3*   h2   u2   p2
Q1  0.38                0.20 0.80 0.72
Q2  0.45       0.24  0.24 0.32 0.68 0.64
Q3  0.25       0.20       0.11 0.89 0.60
Q4  0.30             0.38 0.24 0.76 0.38
Q5  0.48       0.32       0.35 0.65 0.66
Q6  0.50  0.22       0.35 0.43 0.57 0.59
Q7  0.52  0.45            0.48 0.52 0.56
Q8  0.55  0.21 0.28       0.42 0.58 0.71
Q9  0.56       0.37       0.46 0.54 0.67
Q10 0.32       0.34       0.22 0.78 0.46
Q11 0.21       0.32 -0.24 0.21 0.79 0.22
Q12 0.30                  0.14 0.86 0.69
Q13 0.33       0.32       0.21 0.79 0.50
Q14 0.40       0.33       0.28 0.72 0.59
Q15            0.26       0.16 0.84 0.22

With eigenvalues of:
   g  F1*  F2*  F3*
2.40 0.39 0.94 0.47
```

Figure 7

Lertap5 will undertake a DIF analysis using the **Mantel-Haenszel** procedure, another way to contemplate invariance – I would demonstrate it by combining[9] the two xlsx workbooks made for the IRT Rasch analysis, adding a new column with a categorical code of A, for sample Half1, B for sample Half2, and then letting Lertap5 do its thing.

When I had a go at what I've just described, the analysis flagged Q6 as showing DIF – the corresponding output is shown in Figures 8 and 9. No other item was singled out.

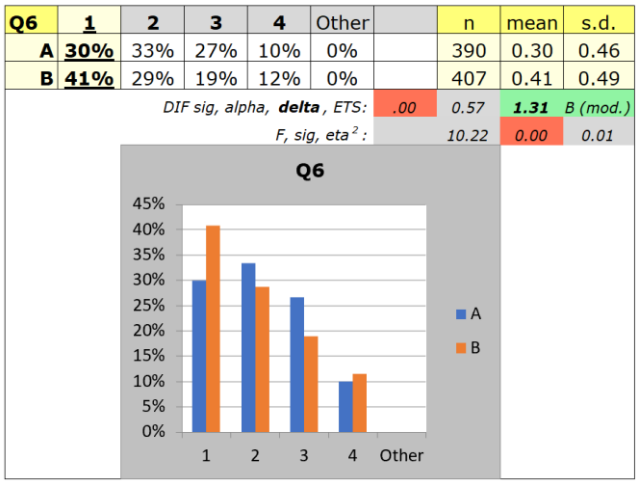

Figure 8

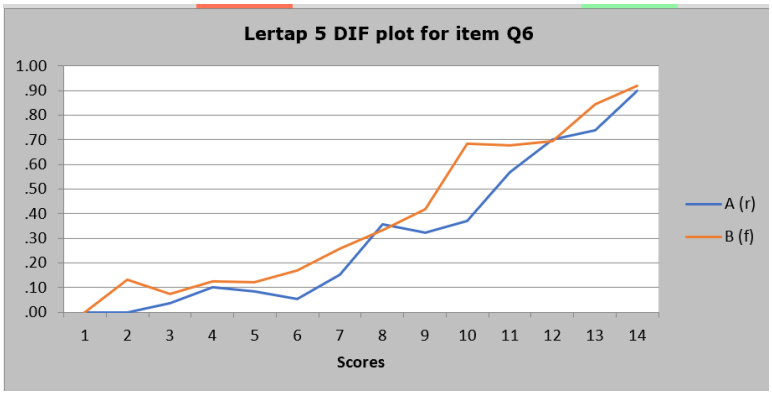

Figure 9

The use of the Mantel-Haenszel procedure might in this case be the most recommended DIF approach, but application of the **difR** package could also be worthwhile. A discussion of using difR with Lertap5 is here.

Possible class exercises

Not all that difficult: the scree plot above is based on the eigenvalues of a product-moment correlation matrix. How might the scree plot change if tetrachoric correlations were used?

Not all that difficult: does the Schmid-Leiman solution vary noticeably over samples, such as Half1 and Half2?

A bit time consuming: if students have been playing along and have two workbooks, such as Half1 and Half2, have them repeat the DIF exercise I demo above. I repeated it my-

[9] Copy the records in one of the workbook's Data sheet and paste them after the last Data record in the other workbook.

self and found Q2 to come to the fore at an ETS B level; Q6 dipped to ETS A level (negligible; it was ETS Level B above, see the green highlighting in Figure 8).

TAM Tutorial 4

Another world beckons those who may like to have their classes explore more measurement topics.

The "FIMS" dataset involves a 14-item cognitive "test" with results from two countries, Australia and Japan, and includes student gender identification. There's gold here; it might be mined in numerous ways.

I would begin by pointing out that the Japan students exhibited stronger performance on many of the items, resulting in higher number-right scores. The quickest way to show this is to use the complete dataset and then make use of the "Breakout score by groups" option in Lertap5. Note that a boxplot may be produced. Easy; possibly enlightening.

Item-level differences may be obtained by using the same dataset and the item responses by groups option. Fairly easy; possibly enlightening.

Such analyses will readily show that the country-level differences are considerable.

Invariance?

Random halves can be obtained quickly from each of the country workbooks, Australia and Japan. I used the "ToHalveAndHold" option within each. Table 6 shows the results from Lertap5, with the Omega1 macro used to obtain the omega values.

Sample	n	mean	s.d.	coefficient alpha	coefficient omega
Japan 1	1,026	8.25 (58.9%)	2.83	0.72	0.75
Japan 2	1,025	8.19 (58.5%)	2.88	0.73	0.76
Australia 1	2,160	6.22 (44.4%)	2.38	0.61	0.65
Australia 2	2,160	6.16 (44.0%)	2.36	0.60	0.63

Table 6

The low alpha and omega values might raise a cautionary flag at the start: if the 14 items are scored as a test, with a possible eye to Rasch IRT, dimensionality could be a question to look at. (Scree plots, as in Figure 6? The Omega1 macro to get a Schmid-Leiman solution, as in Figure 7?)

Within countries, do items have acceptable CTT statistics? They're rather marginal, especially in Australia. A standard Lertap5 analysis will produce the Stats1b scatterplots seen in Figure 10 and Figure 11.

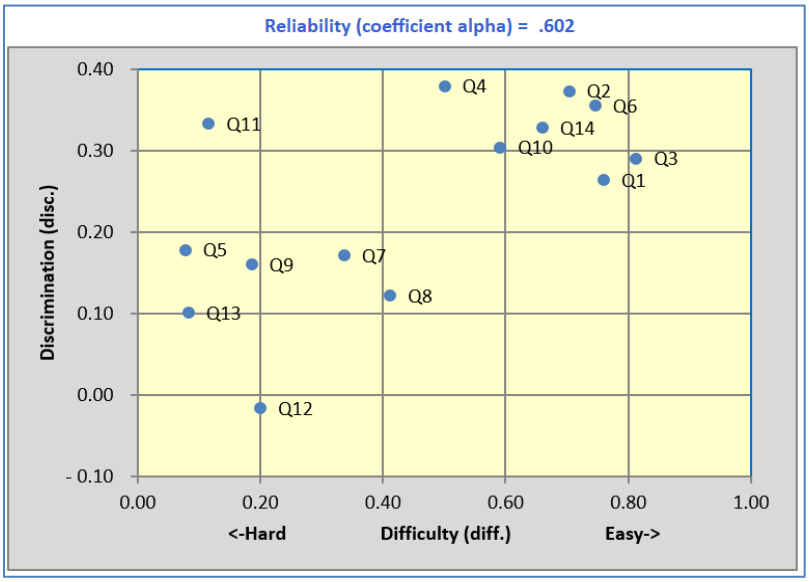

Figure 10: Australia

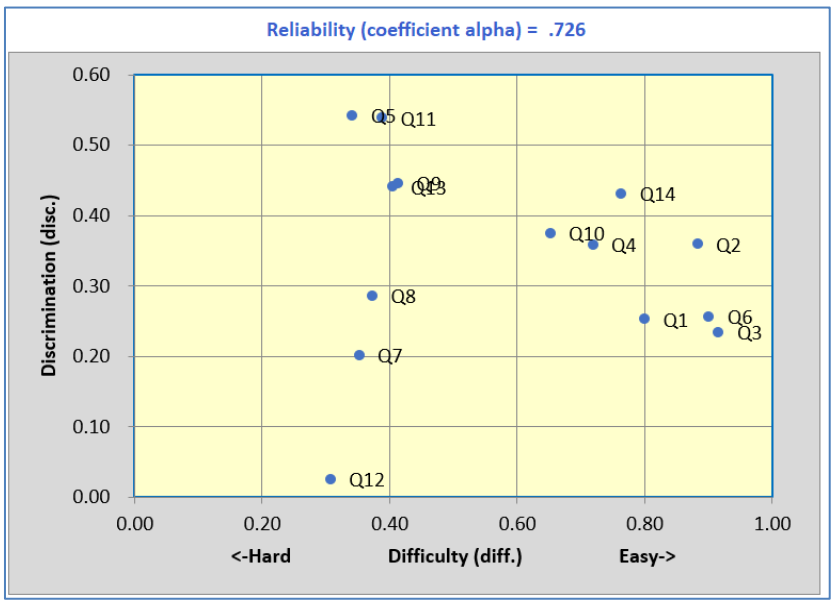

Figure 11: Japan

The items had better discrimination in the Japan sample, with five of the 14 having CTT discrimination above 0.40. In the Australia sample, no item had CTT discrimination above 0.40; five items were below 0.20. Four of the items, Q5, Q9, Q11, and Q13 were harder for the Australia students; Q2, Q3, and Q6 were easier for the Japan students.

Q12 had poor discrimination in both countries. Were this item omitted, coefficient alpha would move up to about 0.63 for Australia, and up to about 0.75 for Japan. These alphas are still on the weak side; number-correct scores will likely have considerable error even after excluding Q12.

Now, concerning invariance per se, the scatterplots turn out to be worth a comment or two. Four of them are quite nice; one, the last, is not so nice but does make quite a worthwhile point.

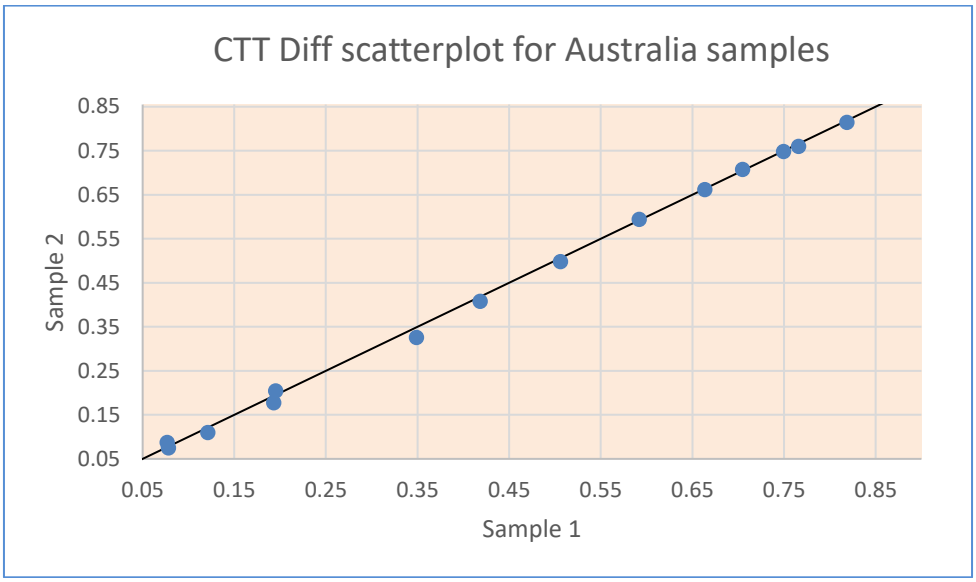

Figure 12: CTT item diffs using Australia samples, r=0.999

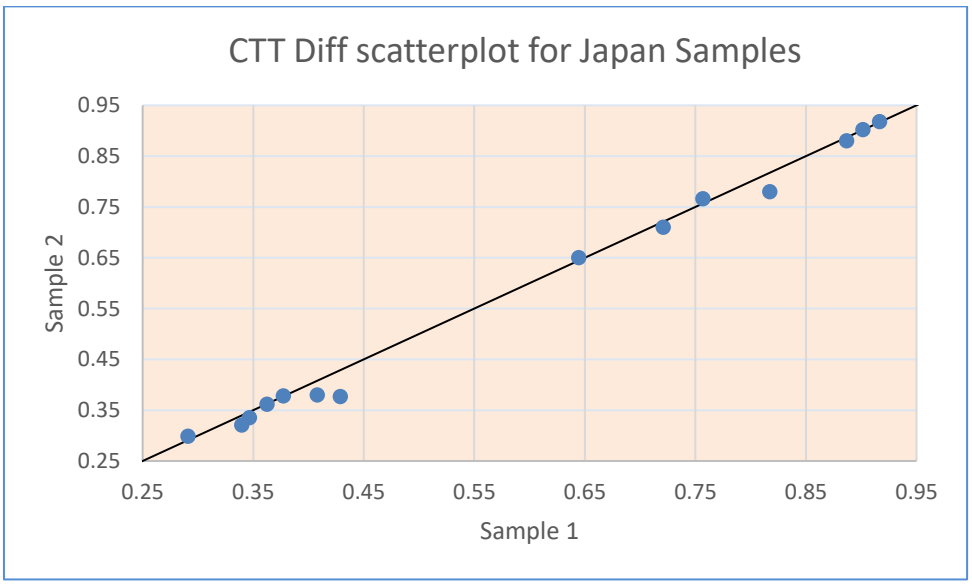

Figure 13: CTT item diffs using Japan samples, r=0.997

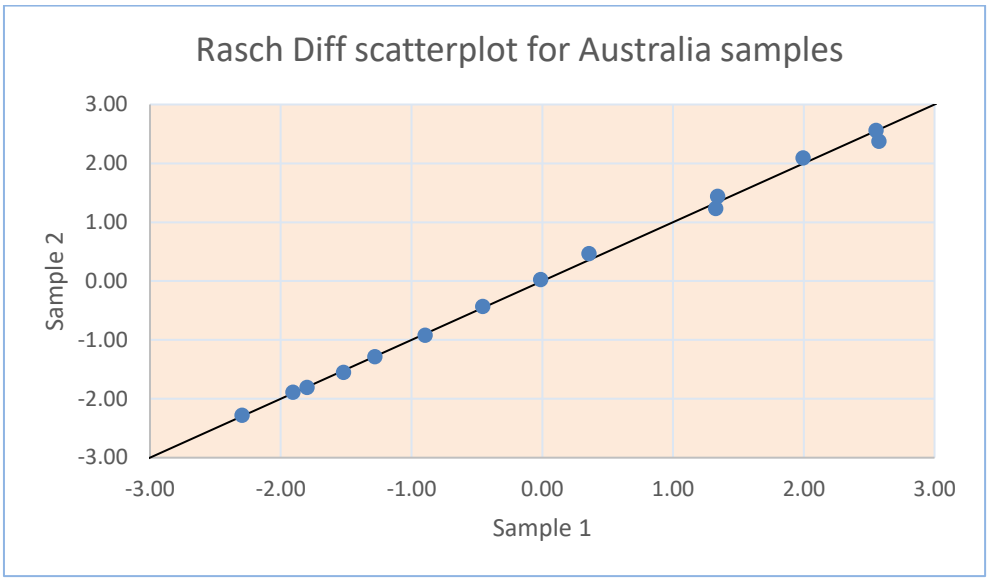

Figure 14: Rasch item diffs using Australia samples, r=.999

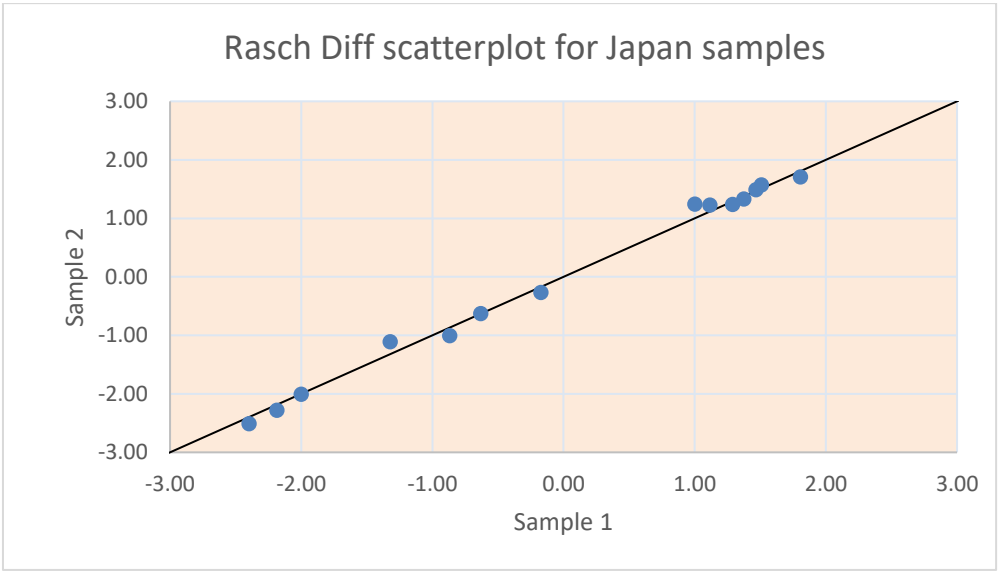

Figure 15: Rasch item diffs using Japan samples, r=.997

As was the case with Tutorial 3 data, CTT and Rasch return highly similar "invariance" pictures in Figures 12 through 15. These plots relate to within-country data samples.

Table 6 and corresponding plot, Figure 16, result from combining samples from both countries, Australia and Japan.

Q1	JPN1	-1.32	Q6	JPN1	-2.18	Q11	JPN1	1.29
Q1	AUS1	-1.91	Q6	AUS1	-1.80	Q11	AUS1	2.00
Q2	JPN1	-2.00	Q7	JPN1	1.47	Q12	JPN1	1.81
Q2	AUS1	-1.52	Q7	AUS1	0.36	Q12	AUS1	1.33
Q3	JPN1	-2.40	Q8	JPN1	1.37	Q13	JPN1	1.12
Q3	AUS1	-2.30	Q8	AUS1	-0.01	Q13	AUS1	2.58
Q4	JPN1	-0.63	Q9	JPN1	1.00	Q14	JPN1	-0.87
Q4	AUS1	-0.46	Q9	AUS1	1.34	Q14	AUS1	-1.28
Q5	JPN1	1.51	Q10	JPN1	-0.17			
Q5	AUS1	2.55	Q10	AUS1	-0.90			

Table 6 (Rasch Diffs, JPN1 and AUS1 samples)

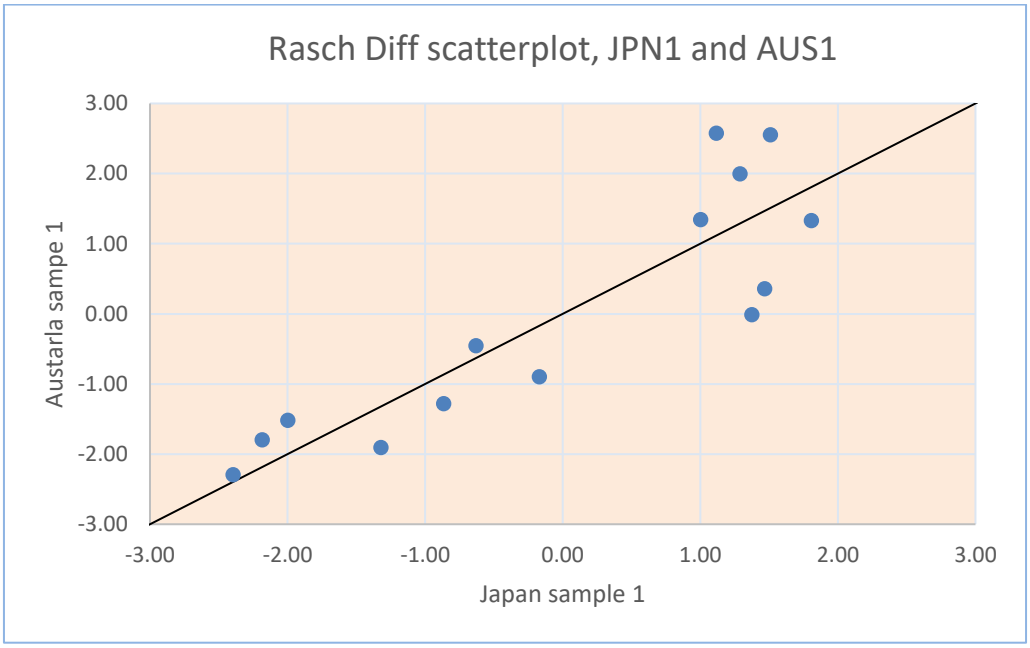

Figure 16: Rasch item diffs using samples from both countries, r=.915

Figure 16 is a cross-country picture plotting the Rasch difficulty estimates from an Australia sample against those from a Japan sample as found in Table 6. Now the "picture" is clearly not quite as clear. There is a pattern for sure; the correlation between the two difficulty estimates is high, but the plot reflects substantially less invariance than seen in the four within-country scatterplots, Figures 12, 13, 14, and 15.

If the Rasch model were to hold this should not be the case. When it holds, Rasch statistics will be sample free (see Wright and Stone (1979).

Kline (2015, p.200) has a discussion about using test items with "clearly distinctive populations", suggesting that, if Rasch methodology is truly sample free, we'd expect Figure 16 to be clearer. Kline goes on to say that "*... 'sample-free' is not an accurate description of Rasch scaling ....*". Best to stick to "*properly stratified samples*". The results obtained in this section appear to lend some support for Kline's suggestion. Table 6, and Figures 10 and 11, indicate that, of the two countries, the Australia sample was considerably weaker. Within countries, Rasch results were invariant, but not so, not to the same extent, when looking at a sample that included students from both countries.

DIF?

The DIF analysis of TAM Tutorial 3 data uncovered an apparent bit of differential item functioning; see Figures 6 and 7. I have not found anything similar with TAM Tutorial 4 data, but, if DIF is an instructional focus, students might look for it using the gender variable in Tutorial 4. This paper has more DIF material with what may serve as better examples.

More IRT exercises?

Look here to extend the use of the RaschAnalysis1 macro. Compare Rasch to the 1PL model by using the EIRT Excel add-in, correlating/scatterplotting its estimates of "b" with Rasch Diffs from the macro (maybe students will be surprised with the outcome[10]).

Look at data fit by getting TAM ICC plots with empirical overlays. Use the work from the University of Western Ontario (UWO) for a more global IRT perspective, with an eye towards model fit; an RMD script linking Lertap5 to UWO work is available here.

References

References are at this webpage.

[10] This is another example of an exercise with a rather minor point behind it. Time permitting it might still be worthwhile, depending on the "level" of the class. Might boost their data manipulation skills if nothing else.