Location: Matrices: Spear Jen Hyp and MAPM folder. Filename: Testing SJ Hyp.doc

Testing the Spearman-Jensen Hypothesis Using the *Items* of the RPM

John Raven

30 Great King St.,
Edinburgh EH3 6QH

Website: www.eyeonsociety.co.uk

Version Date: 31 May 2010

A few weeks ago, I was approached by Jan te Nijenhuis who was looking for further data sets to feed into a meta analysis of studies seeking to test the Spearman-Jensen hypothesis that ethnic (and time) differences are *greater* on those tests that have high ***g*** loadings than on those which depend more obviously on cultural/educational background. More specifically, he was seeking data sets which would enable him to accumulate more studies to test this notion at the *item* level on the RPM. Apparently a number of such studies already exist. The hypothesis is that the ethnic (and time?) differences will be greater on those items which have a high ***g*** loading – ie those which have high item-total score correlations.

This presented me with a number of difficulties … not the least of which is that it has long been known (at least since the time of Guttman) that no "***g*** factor" emerges from inter-correlating and factoring the items of *any* test that satisfies the requirements of Item Response Theory. And Andy Fugard had prepared a dramatic illustration of this by calculating and factoring the item-item correlations of that quintessential "Rasch" scale - the "items" – centimetre marks - of a meter or yard stick such as might be used to measure height (Raven and Fugard, 2008; Fugard, 2008).

If one insists, one can extract "factors" all right … but the interpretation to be placed upon them is quite different from that usually given (Guttman called them "power" factors … factors which group together items of similar difficulty). This methodological error has led to numerous studies purporting to show that the RPM does not measure a single factor but combines several "factors" made up of items measuring "different things" - such as "simple perceptual ability" and "analytic ability". In contrast, IRT analyses had regularly shown that this is not true (at least at the level implied). One simply cannot apply thoughtways derived from Classical Test Theory to measures developed according to the principles of Item Response Theory.

So I wondered what the distribution of item-total score correlations for a meter/yard stick would look like and asked Andy to generate these for the computer-generated data set he had used to calculate the correlations between the items ("centimetre marks") of a 36-item meter/yardstick (which might be used to measure height and assess ethnic differences on this "general factor" of "length").

The results are shown below:

11 11 14 17 20 29 37 41 46 50 54 59 64 68 72 76 78 79 79 79 77 74 70 67 62 56 47 41 36 28 24 23 18 18 15 11

Hmm. Now. What have we got here? The claim is that we are going to be able to test the Spearman-Jensen hypothesis by looking to see whether the ethnic (and time) differences are greatest on those items that have the highest item-total correlations … ie those in the middle of the test.

Somehow it seems that that would not be altogether surprising.

On the other hand, and I am by no means sure about this, I have a feeling that it would be much "easier" to accumulate a "significant difference" in the centre of the distribution than in the tails. It is much easier to increase one's high jumping ability by 1 cm at the lower end of the scale than at the top of the scale (or is it? What makes one think that it is easier for a low ability guy to increase his average high jump by 1 cm than for a really fit guy to do so?)

At this point, I would like to come at the issue from a different angle.

As it happens, a colleague, codenamed "M", recently set out to develop a new set of "Advanced Progressive Matrices" items based on a "non-Carpenter-and-Just" theory of how the items might work. (See also Vodegel Matzen, 1994, and Styles, 2008). In order to distinguish his item set from the classical RAPM Set II set we have here called them MAPM items.

In order to begin to see how successful he had been we persuaded our old friends Anca Dobrean to collect some pilot data and Joerg Prieler to analyse it.

Figure 1.

Chart File name APM_1PL(corrected).doc and xls 31 May 2010

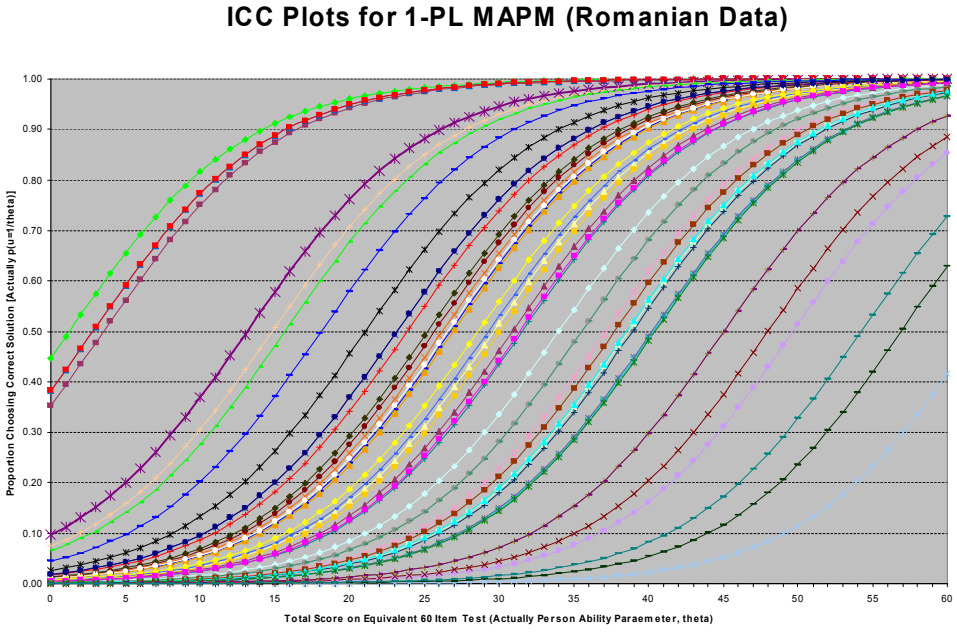


Figure 1 presents a standard 1 PL ("Rasch") analysis of the MAPM data. Note that this is the most common form of IRT analysis conducted by those dabbling in Item Response Theory – researchers who almost invariably do so without undertaking the difficult task of actually

plotting the ICCs (and often, it would seem, without fully understanding the meaning of the item parameters produced by the programs they use).

It looks WOW. Here we have a set of MAPM items which appear to almost achieve the standards required by using a meter/yard stick to measure height.

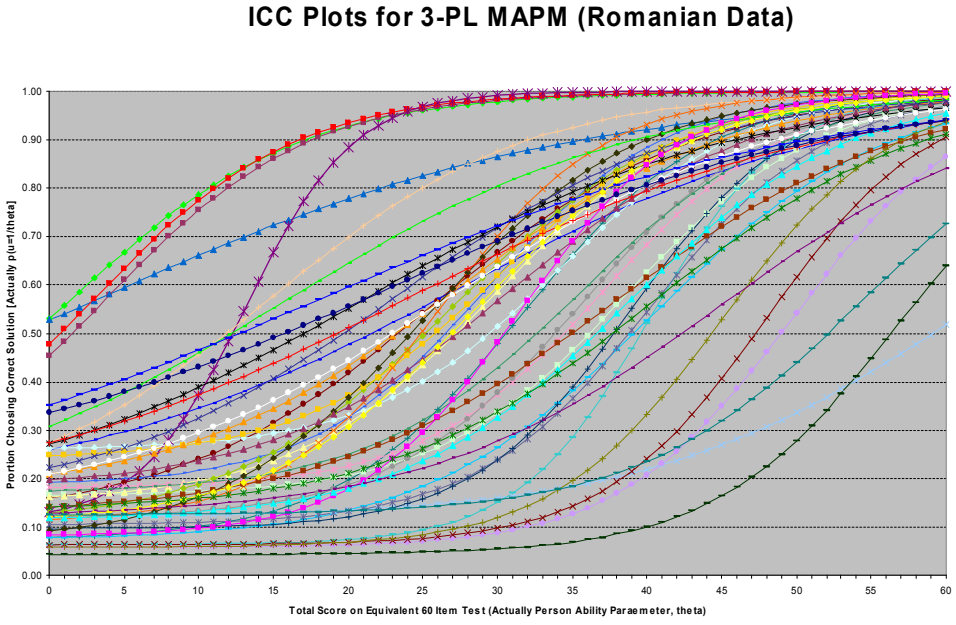
The only additional step would seem to be to eliminate the superfluous items in order to get a set which are equally spaced and thus able to yield a linear Test Characteristic Curve.

Note that in this analysis all items appear to have similar discriminatory power (= item-total score correlations). I am not really sure what these would look like if calculated as Pearson Product Moment correlations. But I strongly suspect that the items in the centre of the distribution would have relatively high item-total correlations because they are discriminating across the full range of ability while those in the tails would have lower correlations because they are discriminating only within a narrow range of ability. So any substantive conclusions based on these differential correlations would be in danger of being misleading.

But now let us have a look at Figure 2 – a "3 PL" analysis. This, as Raven, Prieler, & Benesch (2008) (well, actually, it was Gerhard Fisher, but he is a bit shy) have shown, comes closer to revealing what the "empirical curves" (Raven 1938) *really* show. (3 PL ICCs provide indices of difficulty, discriminative power (slope, correlation), and "guessing" [proportion of respondents getting the item "right" without apparently having the level of ability required to do so].)

Figure 2

Chart File name APM_3PL(corrected).doc and xls 31 May 2010



Now we see (approximately) what is *really* going on!

There are indeed some items that have "good" ICCs. Few people who lack the requisite ability get them right, the curves rise steeply (ie they have good discriminative power), and few of those with high ability ever get them wrong.

But there are far too many items that have lousy ICCs. Far too many "low ability" people get them right (by chance???) and many high ability people get them wrong.

Now, what is likely to emerge if one starts looking at ethnic and time differences on these items?

It seems to me unlikely that one is going to get large ethnic or time differences on them. But this will not be because they are not good measures of whatever it is that the test is measuring … perhaps "psychometric ***g***", perhaps "meaning making ability" … (and one will get no help in trying to answer this question by inter-correlating and factoring the items!) but because they are lousy measures of anything.

But what of the items that have good ICCs in the tails of the distribution?

If I am right in what I said above, these are going to have low Pearson item-total score correlations. But they are good items *at that level of ability.* As Prieler and I (2008) (but really Fischer!) have shown, differences between groups and over time depend on the absolute level of difficulty of the test being used relative to the population studied and the shape of the Test Characteristic Curve. Here we are taking the TCC back to the item level – the ICC. They are not going to show differences between our ethnic groups, not because they have "low g loadings", but because they are far from the mean of the distribution where there are few people and it is "difficult to make a difference".

In reality, the pattern of item-total score correlations is even more of a mess than even I had suspected. Joerg has calculated these using a standard Classical Test Theory package and the results are shown below. However, to understand the print out one first has to know that the 60 item test shown on the x axes in the Figures above came from data assembled from combining two sets of 30 overlapping items which had been prepared to avoid overloading those taking the tests.

Item-Scale-Statistics for Form A			Item-Scale-Statistics for Form B		
Item	Item-Scale-Correlation	Cronbachs Alpha, if Item deleted	Item	Item-Scale-Correlation	Cronbachs Alpha, if Item deleted
ITEM01	,245	,636	ITEM01	,135	,563
ITEM02	,292	,629	ITEM02	,221	,551
ITEM03	,214	,638	ITEM03	,267	,552
ITEM04	,120	,647	ITEM04	,109	,567
ITEM05	,080	,648	ITEM05	,118	,564
ITEM06	,347	,627	ITEM06	,217	,556
ITEM07	,363	,621	ITEM07	,411	,533
ITEM08	,317	,630	ITEM08	,210	,553
ITEM09	,363	,621	ITEM09	,072	,572
ITEM10	-,051	,664	ITEM10	,347	,532
ITEM11	,276	,631	ITEM11	,168	,558
ITEM12	,213	,638	ITEM12	,472	,515
ITEM13	,548	,602	ITEM13	-,030	,584
ITEM14	,368	,620	ITEM14	,093	,569
ITEM15	,090	,650	ITEM15	,367	,530
ITEM16	,250	,635	ITEM16	,106	,566
ITEM17	,171	,641	ITEM17	,485	,513
ITEM18	,221	,637	ITEM18	,018	,577
ITEM19	-,008	,659	ITEM19	,070	,570
ITEM20	,269	,633	ITEM20	-,228	,602
ITEM21	,026	,654	ITEM21	,321	,546
ITEM22	,141	,644	ITEM22	,295	,540
ITEM23	,205	,642	ITEM23	,130	,563
ITEM24	,157	,643	ITEM24	-,048	,572
ITEM25	,156	,643	ITEM25	-,073	,591
ITEM26	,128	,645	ITEM26	,389	,537
ITEM27	-,116	,665	ITEM27	-,030	,582
ITEM28	,074	,650	ITEM28	-,240	,598
ITEM29	,128	,644	ITEM29	,195	,557
ITEM30	,235	,635	ITEM30	,099	,566

Reliability				
Form A			Form B	
Cronbachs Alpha	Amount of Items		Cronbachs Alpha	Amount of Items
,647	30		,569	30

So, where does all this leave those, like Jan, who have sought to test the Spearman-Jensen hypothesis at the item level?

References

Fugard, A. (2008, April 14) Comparing IRT and EFA on Raven's like tests. Edinburgh Psychology R-users wiki: getting things done with R. Retrieved May 28, 2010, from http://psy-ed.wikidot.com/ravenslike-irt-fa

Prieler, J. & Raven, J. (2008). Problems in the measurement of change (with particular reference to individual change [gain] scores) and their potential solution using IRT. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics*. Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 7, pp. 173-210) or, earlier version: *WebPsychEmpiricis*t http://www.wpe.info/papers_table.html http://home.earthlink.net/~rkmc/vault/priravf/prirav.pdf

Raven, J., Prieler, J., & Benesch, M. (2008) Using the Romanian data to replicate the IRT-based item analysis of the SPM+: Striking achievements, pitfalls, and lessons. In: J. Raven & J. Raven (Eds.), *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metric*s (Chapter 5). Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000. Also available at: http://www.wpe.info/papers_table.html

Raven, J., & Fugard, A. (2008). What's wrong with factor-analyzing tests conforming to the requirements of item response theory? *WebPsychEmpiricist,* May 23. http://wpe.info/papers_table.html or http://eyeonsociety.co.uk/resources/fairttsts.pdf

Styles, I. (2008). Linking psychometric and cognitive-developmental frameworks for thinking about intellectual functioning. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics*. Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 2, pp. 69-98). Also: *WebPsychEmpiricist*. http://wpe.info/papers_table.html

Vodegel-Matzen, L. B. L. (1994). *Performance on Raven's Progressive Matrices*. Ph.D. Thesis, University of Amsterdam.