the WINNOWER

SOCIAL SCIENCES

Spearman's hypothesis on item-level data from Raven's Standard Progressive Matrices: A replication and extension

EMIL O. W. KIRKEGAARD

CORRESPONDENCE:
emil@emilkirkegaard.dk

DATE RECEIVED:
June 10, 2015

Abstract

Item-level data from Raven's Standard Progressive Matrices was compiled for 12 diverse groups from previously published studies. The method of correlated vectors was used on every possible pair of groups with available data (45 comparisons). Depending on exact method chosen, the mean and mean MCV correlation was about .46/51. Only 2/1 of 45 were negative. Spearman's hypothesis is confirmed for item-level data from the Standard Progressive Matrices.

Introduction and method

The method of correlated vectors (MCV) is a statistical method invented by Arthur Jensen (1998, p. 371, see also appendix B). The purpose of it is to measure to which degree a latent variable is responsible for an observed correlation between an aggregate measure and a criteria variable. Jensen had in mind the general factor of cognitive ability data (the *g* factor) as measured by various IQ tests and their subtests, and criteria variables such as brain size, however the method is applicable to any latent trait (e.g. general socioeconomic factor Kirkegaard, 2014a, b). When this method is applied to group differences, particularly ethnoracial ones, it is called *Spearman's hypothesis* (SH) because Spearman was the first to note it in his 1927 book.

By now, several large studies and meta-analysis of MCV results for group differences have been published (te Nijenhuis et al (2015a, 2015b, 2014), Jensen (1985)). These studies generally support the hypothesis. Almost all studies use subtest loadings instead of item loadings. This is probably because psychologists are reluctant to share their data (Wicherts, et al, 2006) and as a result there are few open datasets available to use for this purpose. Furthermore, before the introduction of modern computers and the internet, it was impractical to share item-level data. There are advantages and disadvantages to using item-level data over subtest-level data. There are more items than subtests which means that the vectors will be longer and thus sampling error will be smaller. On the other hand, items are less reliable and less pure measures of the g factor which introduces both error and more non-g ability variance.

The recent study by Nijenhuis et al (2015a) however, employed item-level data from Raven's Standard Progressive Matrices (SPM) and included a diverse set of samples (Libyan, Russian, South African, Roma from Serbia, Moroccan and Spanish). The authors did not use their collected data to its full extent, presumably because they were comparing the groups (semi-)manually. To compare all

combinations with a dataset of e.g. 10 groups means that one has to do 45 comparisons (10*9/2). However, this task can easily be overcome with programming skills, and I thus saw an opportunity to gather more data regarding SH.

The authors did not provide the data in the paper despite it being easy to include it in tables. However, the data was available from the primary studies they cited in most cases. Thus, I collected the data from their data sources (it can be found in the supplementary material). This resulted in data from 12 samples of which 10 had both difficulty and item-whole correlations data. Table 1 gives an overview of the datasets:

Table 1- Overview of samples

Short name	Race	Selection	N	Year	Ref	Country	Description
A1	African	Undergraduates	173	2000	Rushton and Skuy 2000	South Africa	University of the Witwatersrand and the Rand Afrikaans University in Johannesburg, South Africa
W1	European	Undergraduates	136	2000	Rushton and Skuy 2000	South Africa	University of the Witwatersrand and the Rand Afrikaans University in Johannesburg, South Africa
W2	European	Std 7 classes	1056	1992	Owen 1992	South Africa	20 schools in the Pretoria-Witwatersrand-Vereeniging (PWV) area and 10 schools in the Cape Peninsula
C1	Colored (African European)	Std 7 classes	778	1992	Owen 1992	South Africa	20 coloured schools in the Cape Peninsula
I1	Indian	Std 7 classes	1063	1992	Owen 1992	South Africa	30 schools selected at random from the list of high schools in and around Durban
A2	African	Std 7 classes	1093	1992	Owen 1992	South Africa	Three schools in the PWV area and 25 schools in KwaZulu (Natal)
A3	African	First year Engineering students	198	2002	Rushton et al 2002	South Africa	First-year students from the Faculties of Engineering and the Built Environment at the University of the Witwatersrand
I2	Indian	First year Engineering students	58	2002	Rushton et al 2002	South Africa	First-year students from the Faculties of Engineering and the Built Environment at the University of the Witwatersrand
W3	European	First year Engineering students	86	2002	Rushton et al 2002	South Africa	First-year students from the Faculties of Engineering and the Built Environment at the University of the Witwatersrand
R1	Roma	Adults ages 16 to 66	231	2004.5	Rushton et al 2007	Serbia	The communities (i.e., Drenovac, Mirijevo, and Rakovica) are in the vicinity of Belgrade
W4	European	Adults ages 18 to 65	258	2012	Diaz et al 2012	Spain	Mainly from the city of Valencia
NA1	North African	Adults ages 18 to 50	202	2012	Diaz et al 2012	Morocco	Casablanca, Marrakech, Meknes and Tangiers

Item-whole correlations and item loadings

The data in the papers did usually not contain the actual factor loadings of the items. Instead, they contained the item-whole correlations. The authors argue that one can use these because of the high correlation of unweighted means with extracted g-factors (often, r=.99, e.g. Kirkegaard, in review). Some studies did provide both loadings and item-whole correlations, yet the authors did not correlate them to see how good proxies the item-whole correlations are for the loadings. I calculated this for the 4 studies that included both metrics. Results are shown in Table 2.

Table 2 – Item-whole correlations x g-loadings in 4 studies.

	W2	C1	I1	A2
W2	*0.549*	0.099	0.327	0.197
C1	0.695	*0.900*	0.843	0.920
I1	0.616	0.591	*0.782*	0.686
A2	0.626	0.882	0.799	*0.981*

Note: Within sample correlations between item-whole correlations and item factor loadings are in the diagonal, marked with italic.

As can be seen, the item-whole correlations were not in all cases great proxies for the actual loadings.

To further test this idea, I calculated the item-whole correlations and the factor loadings (first factor, minimum residuals) in the open Wicherts dataset (N=500ish, Dutch university students, see Wicherts and Bakker 2012) tested on Raven's Advanced Progressive Matrices. The correlation was .89. Thus, aside from the odd result in the W2 sample, item-whole correlations were a reasonable proxy for the factor loadings.

Item difficulties across samples

If two groups are tested on the same test and this test measures the same trait in both groups, then even if the groups have different mean trait levels, the order of difficulty of the items or subtests should be similar. Rushton et al (2000, 2002, 2007) have examined this in previous studies and found it generally to be the case. Table 3 below shows the cross-sample correlations of item difficulties.

Table 3 – Intercorrelations between item difficulties in 12 samples

	A1	W1	W2	C1	I1	A2	A3	I2	W3	R1	NA1	W4
A1	1	0.88	0.98	0.96	0.99	0.86	0.96	0.89	0.79	0.89	0.95	0.93
W1	0.881		0.93	0.79	0.87	0.65	0.96	0.97	0.94	0.7		0.92 0.95
W2	0.980.931			0.95	0.98	0.82	0.97	0.92	0.84	0.86	0.96	0.95
C1	0.960.790.951				0.98	0.94	0.89	0.81	0.69	0.95	0.92	0.87
I1	0.990.870.980.981					0.88	0.95	0.88	0.79	0.91	0.95	0.92
A2	0.860.650.820.940.881						0.76	0.68	0.56	0.97	0.82	0.76
A3	0.960.960.970.890.950.761							0.96	0.9	0.8	0.95	0.96
I2	0.890.970.920.810.880.680.961								0.92	0.72	0.91	0.92
W3	0.790.940.840.690.790.560.9							0.921		0.6	0.88	0.91
R1	0.890.7	0.86	0.95	0.91	0.97	0.8	0.72	0.6	1		0.86	0.8
NA1	0.950.920.960.920.950.820.950.910.880.861											0.97
W4	0.930.950.950.870.920.760.960.920.910.8										0.97	1

The mean intercorrelation is.88. This is quite remarkable given the diversity of the samples.

Item-whole correlations across samples

Given the above, one might expect similar results for the item-whole correlations. This however is not so. Results are in Table 4.

Table 4 – Intercorrelations between item-whole correlations in 10 samples

	A1	W1	W2	C1	I1	A2	A3	I2	W3	R1
A1	1	-0.2	0.59	0.58	0.73	0.54	0.27	0.04	-0.3	0.57
W1	-0.2	1	0.17	-0.59	-0.25	-0.68	0.42	0.51	0.55	-0.55
W2	0.59	0.17	1	0.44	0.79	0.29	0.61	0.25	0.02	0.39
C1	0.58	-0.59	0.44	1	0.79	0.94	0.01	-0.25	-0.49	0.78
I1	0.73	-0.25	0.79	0.79	1	0.69	0.42	0.09	-0.33	0.63
A2	0.54	-0.68	0.29	0.94	0.69	1	-0.13	-0.3	-0.52	0.77
A3	0.27	0.42	0.61	0.01	0.42	-0.13	1	0.26	0.37	0.02
I2	0.04	0.51	0.25	-0.25	0.09	-0.3	0.26	1	0.34	-0.21
W3	-0.3	0.55	0.02	-0.49	-0.33	-0.52	0.37	0.34	1	-0.49
R1	0.57	-0.55	0.39	0.78	0.63	0.77	0.02	-0.21	-0.49	1

Note: The last two samples, NA1 and W4, did not have item-whole correlation data.

The reason for this state of affairs is that the factor loadings change when the group mean trait level changes. For many samples, most of the items were too easy (passing rates at or very close to 100%). When there is no variation in a variable, one cannot calculate a correlation to some other variable. This means that for a substantial number of items for multiple samples, there was missing data for the items.

The lack of cross-sample consistency in item-whole correlations may also explain the weak MCV results in Diaz et al, 2012 since they used g-loadings from another study instead of from their own samples.

Spearman's hypothesis using one static vector of estimated factor loadings

Some of the sample had rather low sample sizes (I2, N=58, W3, N=86). Thus one might get the idea to use the item-whole correlations from one or more of the large samples for comparisons involving other groups. In fact, given the instability of item-whole correlations across sample as can be seen in Table 4, this is a bad idea. However, for sake of completeness, I calculated the results based on this anyway. As the best estimate of factor loadings, I averaged the item-whole correlations data from the four largest samples (W2, C1, I1 and A2).

Using this vector of item-whole correlations, I used MCV on every possible sample comparison. Because there were 12 samples, this number is 66. MCV analysis was done by subtracting the lower scoring sample's item difficulties from the higher scoring sample's thus producing a vector of the sample difference on each item. This vector I correlated with the vector of item-whole correlations. The results are shown in Table 5.

Table 5 – MCV correlations of group differences across 12 samples using 1 static item-whole correlations

	A1	W1	W2	C1	I1	A2	A3	I2	W3	R1	NA1	W4
A1	NA	-0.15	0.2	0.42	0.1	0.83	-0.03	-0.12	-0.26	0.8	-0.35	-0.32
W1	-0.15	NA	-0.31	0.07	-0.14	0.47	-0.29	-0.19	-0.4	0.4	0.31	0.06
W2	0.2	-0.31	NA	0.56	0.4	0.86	-0.27	-0.28	-0.38	0.83	-0.35	-0.46
C1	0.42	0.07	0.56	NA	0.53	0.88	0.23	0.11	-0.06	0.64	-0.23	-0.05
I1	0.1	-0.14	0.4	0.53	NA	0.88	-0.02	-0.1	-0.24	0.83	-0.45	-0.29
A2	0.83	0.47	0.86	0.88	0.88	NA	0.66	0.52	0.32	0.2	0.42	0.4
A3	-0.03	-0.29	-0.27	0.23	-0.02	0.66	NA	-0.17	-0.41	0.61	0.43	-0.43
I2	-0.12	-0.19	-0.28	0.11	-0.1	0.52	-0.17	NA	-0.37	0.46	0.33	-0.25
W3	-0.26	-0.4	-0.38	-0.06	-0.24	0.32	-0.41	-0.37	NA	0.23	-0.05	-0.11
R1	0.8	0.4	0.83	0.64	0.83	0.2	0.61	0.46	0.23	NA	0.36	0.32
NA1	-0.35	0.31	-0.35	-0.23	-0.45	0.42	0.43	0.33	-0.05	0.36	NA	-0.03
W4	-0.32	0.06	-0.46	-0.05	-0.29	0.4	-0.43	-0.25	-0.11	0.32	-0.03	NA

As one can see, the results are all over the place. The mean MCV correlation is .12.

Spearman's hypothesis using a variable vector of estimated factor loadings

Since item-whole correlations varied from sample to sample, another idea is to use the samples' item-whole correlations. I used the unweighted mean of the item-whole correlations for each item (te Nijenhuis et al used a weighted mean). In some cases, only one sample has item-whole correlations for some items (because the other sample had no variance on the item, i.e. 100% get it right). In these cases, one can choose to use the value from the remaining sample, or one can ignore the item and calculated MCV based on the remaining items. I calculated results using both methods, they are shown in Table 6 and 7.

Table 6 – MCV correlations of group differences across 10 samples using variable item-whole correlations, method 1

	A1	W1	W2	C1	I1	A2	A3	I2	W3	R1
A1	NA	0.79	0.39	0.29	0.05	0.7	0.48	0.41	0.37	0.71
W1	0.79	NA	0.8	0.5	0.76	0.51	0.79	0.4	0.6	0.54
W2	0.39	0.8	NA	0.68	0.63	0.85	0.43	0.47	0.5	0.79
C1	0.29	0.5	0.68	NA	0.52	0.88	0.32	0.3	-0.09	0.67
I1	0.05	0.76	0.63	0.52	NA	0.84	0.47	0.4	0.31	0.79
A2	0.7	0.51	0.85	0.88	0.84	NA	0.57	0.43	-0.03	0.22
A3	0.48	0.79	0.43	0.32	0.47	0.57	NA	0.38	0.66	0.64
I2	0.41	0.4	0.47	0.3	0.4	0.43	0.38	NA	0.6	0.44
W3	0.37	0.6	0.5	-0.09	0.31	-0.03	0.66	0.6	NA	0.2
R1	0.71	0.54	0.79	0.67	0.79	0.22	0.64	0.44	0.2	NA

Table 6 – MCV correlations of group differences across 10 samples using variable item-whole correlations, method 2

	A1	W1	W2	C1	I1	A2	A3	I2	W3	R1
A1	NA	0.42	0.4	0.33	0.06	0.72	0.48	0.3	0.15	0.74
W1	0.42	NA	0.72	0.14	0.52	0.18	0.7	0.44	0.65	0.35
W2	0.4	0.72	NA	0.68	0.63	0.85	0.44	0.53	0.5	0.79
C1	0.33	0.14	0.68	NA	0.52	0.88	0.39	0.19	-0.08	0.67
I1	0.06	0.52	0.63	0.52	NA	0.84	0.51	0.4	0.28	0.79
A2	0.72	0.18	0.85	0.88	0.84	NA	0.62	0.3	0.02	0.22
A3	0.48	0.7	0.44	0.39	0.51	0.62	NA	0.42	0.55	0.67
I2	0.3	0.44	0.53	0.19	0.4	0.3	0.42	NA	0.58	0.35
W3	0.15	0.65	0.5	-0.08	0.28	0.02	0.55	0.58	NA	0.08
R1	0.74	0.35	0.79	0.67	0.79	0.22	0.67	0.35	0.08	NA

Nearly all results are positive using either method. The results are slightly stronger when ignoring items where both samples do not have item-whole correlation data. A better way to visualize the results is to use a histogram with a density curve inputted, as shown in Figure 1 and 2.

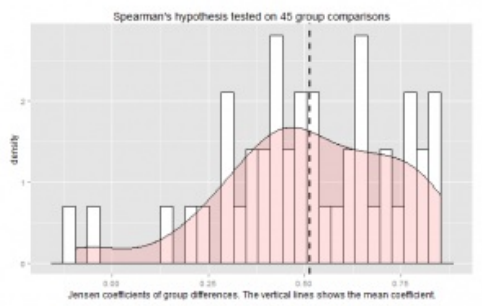

Figure 1 – Histogram of MCV results using method 1

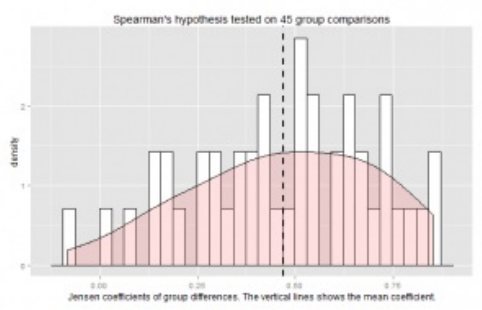

Figure 2 – Histogram of MCV results using method 2

Note: The vertical line shows the mean value.

The mean/median result for method 1 was .51/.50, and .46/.48 for method 2. Almost all MCV results were positive, there were only 2/45 that were negative for method 1, and 1/45 for method 2.

Mean MCV value by sample and moderator analysis

It is interesting to examine the mean MCV value by sample. They are shown in Table 7.

Table 7 – MCV correlation means, SDs, and medians by sample

Sample	mean	SD	median
A1	0.46	0.24	0.41
W1	0.63	0.15	0.60
W2	0.62	0.18	0.63
C1	0.45	0.29	0.50
I1	0.53	0.26	0.52
A2	0.55	0.31	0.57
A3	0.53	0.15	0.48
I2	0.43	0.08	0.41
W3	0.35	0.28	0.37
R1	0.56	0.23	0.64

There is no obvious racial pattern. Instead, one might expect the relatively lower result of some samples to be due to sampling error. MCV is extra sensitive to sampling error. If so, the mean correlation should be higher for the larger samples. To see if this was the case, I calculated the rank-order correlation between sample size and sample mean MCV, r=.45/.65 using method 1 or 2 respectively. Rank-order was used because the effect of sample size on sampling error is non-linear. Figure 3 shows the scatter plot of this.

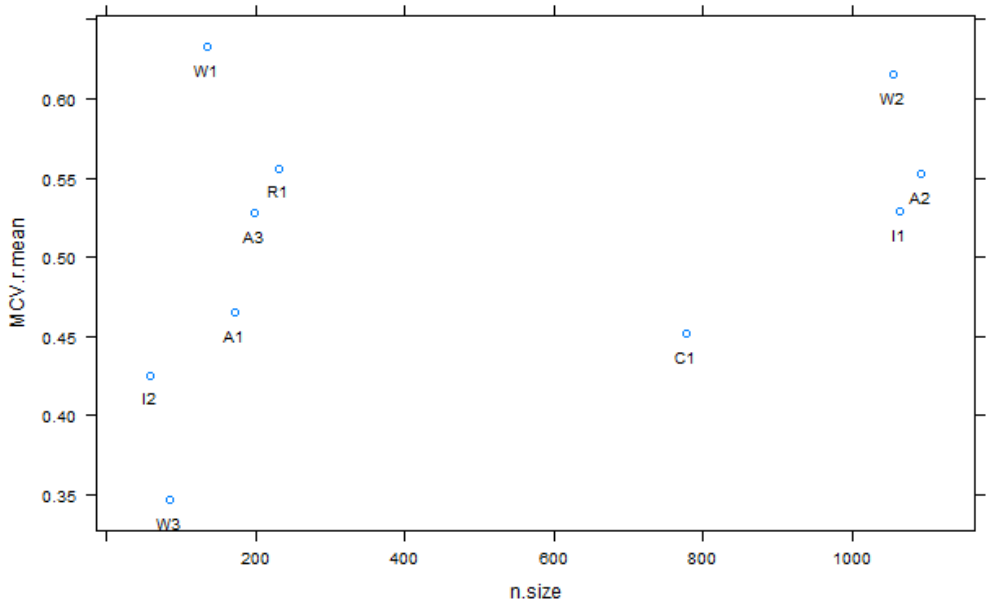

Figure 3 – Sample size as a moderator variable at the sample mean-level

One can also examine sample size as a moderating variable as the comparison-level. This increases the number of datapoints to 45. I used the harmonic mean of the 2 samples as the sample size metric. Figure 4 shows a scatter plot of this.

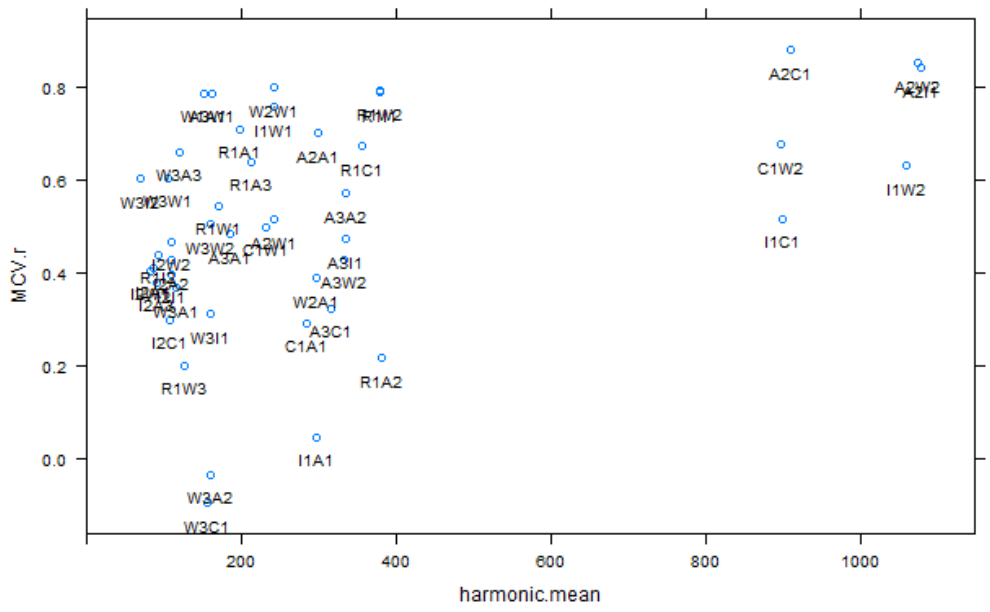

Figure 4 – Sample size as a moderator variable at the comparison-level

The rank-order correlations are .45/.44 using method 1/2 data. We can see in the plot that the results from the 6 largest comparisons (harmonic mean sample size>800) range from .52 to .88 with a mean of .74/.73 and SD of .15 using method 1/2 results. For the smaller studies (harmonic mean sample size<800), the results range from -.09/-.08 to .89/.79 with a mean of .48/.43 and SD of .22/.23 using method 1/2 results. The results from the smaller studies vary more, as expected with their higher

sampling error.

I also examine the group difference size as a moderator variable. I computed this as the difference between the mean item difficulty by the groups. However, it had a near-zero relationship to the MCV results (rank-order r=.03, method 1 data).

Discussion and conclusion

Spearman's hypothesis has been decisively confirmed using item-level data from Raven's Standard Progressive Matrices. The analysis presented here can easily be extended to cover more datasets, as well as item-level data from other IQ tests. Researchers should compile such data into open datasets so they can be used for future studies.

It is interesting to note the consistency of results within and across samples that differ in race. Race differences in general intelligence as measured by the SPM appear to be just like those within races.

Supplementary material

R code and dataset is available at the Open Science Framework repository.

References

- Dıaz, A., & Sellami, K. Infanzó n, E., Lanzó n, T., & Lynn, R.(2012). A comparative study of general intelligence in Spanish and Moroccan samples. Spanish Journal of Psychology, 15(2), 526-532.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. Behavioral and Brain Sciences, 8(02), 193-219.
- Kirkegaard, E. O. (2014a). The international general socioeconomic factor: Factor analyzing international rankings. *Open Differential Psychology*.
- Kirkegaard, E. O. (2014b). Crime, income, educational attainment and employment among immigrant groups in Norway and Finland. *Open Differential Psychology*.
- Kirkegaard, E. O. W. (in review). Examining the ICAR and CRT tests in a Danish student sample. *Open Differential Psychology.*
- Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. Personality and Individual Differences, 13(2), 149-159.
- Rushton, J. P., Čvorović, J., & Bons, T. A. (2007). General mental ability in South Asians: Data from three Roma (Gypsy) Communities in Serbia. Intelligence, 35, 1-12.
- Rushton, J. P., Skuy, M., & Fridjhon, P. (2002). Jensen effects among African, Indian, and White engineering students in South Africa on Raven's Standard Progressive Matrices. Intelligence, 30, 409-423.
- Rushton, J. P., & Skuy, M. (2000). Performance on Raven's Matrices by African and White university students in South Africa. Intelligence, 28, 251-265.
- Spearman, C. (1927). The abilities of man.
- te Nijenhuis, J., Al-Shahomee, A. A., van den Hoek, M., Grigoriev, A., and Repko, J. (2015a). Spearman's hypothesis tested comparing Libyan adults with various other groups of adults on the items of the Standard Progressive Matrices. *Intelligence*. Volume 50, May–June 2015, Pages 114–117
- te Nijenhuis, J., David, H., Metzen, D., & Armstrong, E. L. (2014). Spearman's hypothesis tested on European Jews vs non-Jewish Whites and vs Oriental Jews: Two meta-analyses. Intelligence, 44, 15-18.
- te Nijenhuis, J., van den Hoek, M., & Armstrong, E. L. (2015b). Spearman's hypothesis and Amerindians: A meta-analysis. Intelligence, 50, 87-92.
- Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726.
- Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish

your data too?. Intelligence, 40(2), 73-76.