International Journal of Educational and Psychological Studies – Vol. 3, No. 3, 2018, pp. 720 - 726
e-ISSN 2520-4149 , p-ISSN 2520-4130
Available online at http:// www.refaad.com

Spearman's hypothesis tested in Yemen on the items of the Standard Progressive Matrices Plus: A reply to Díaz, Sellami, Infanzón, Lanzón, & Lynn (2012)

Salaheldin Farah Attallah Bakhiet[1,*], Jan te Nijenhuis[2,3], Michael van den Hoek[2] Mohammed Mohammed Ateik Al-Khadher[1],and Shibaev Vladimir[4]

[1]King Saud University, Riyadh, Saudi Arabia
[2]Work and Organizational Psychology, University of Amsterdam, the Netherlands
[3]National Research Center for Dementia, Chosun University, Gwangju, Korea
[4]Far Eastern Federal University, Vladivostok, Russia
[*]Correspondence Author: Salaheldin Farrah Attallah Bakhiet, King Saud University, College of Education, Department of Special Education, KSA
E-mail: sLh9999@yahoo.com

Abstract:

Spearman's hypothesis simply states that differences between groups on an IQ test are a function of the general intelligence (*g*). At the item level this would mean the magnitude of the differences between groups are smaller on low-*g*-loading items and larger on items with a high *g* loading. An empirical test by Díaz, Sellami, Infanzón, Lanzón, & Lynn (2012) comparing Spanish and Moroccans taking the Raven's Progressive Matrices showed no support for Spearman's hypothesis, whereas other empirical studies showed modest to strong support. In the present study we tested whether another sample from the Arab world would replicate the outcomes of Díaz et al. of no support for Spearman's hypothesis by computing a correlation between *g* loadings and group differences on items of an IQ tests and checking whether it showed a negative correlation. Children from Yemen (total $N = 1916$) were compared with a group of Russian children ($N = 426$) and a group of Kazakh children ($N = 656$) on the Standard Progressive Matrices Plus yielding outcomes of Spearman's hypothesis of, respectively, $r = .86$ ($p < .0005$), and $r = .76$ ($p < .0005$). The $N_{harmonic}$-weighted average correlation of .80 has a credibility interval ranging from .74 to .86. Based on the results we conclude that Spearman's hypothesis holds even at the item level and that the Díaz et al. study appears to be an outlier.

Keywords: Spearman's hypothesis, Yemen, IQ, *g* loadings, group differences.

Highlights
Spearman's hypothesis: group differences are a function of *g* loadings
Spearman's hypothesis is tested comparing children from Yemen, Russia, and Kazakhstan
Spearman's hypothesis at the level of items of the SPM+ is strongly confirmed with mean $r = .80$.

Introduction:

Large-scale research has shown that there are large differences in the mean total scores of Blacks and Whites on traditional IQ tests and that there is no evidence that there is a cultural bias towards Blacks in these tests that would disadvantage Blacks. (Jensen, 1980). The mean Black/White differences vary across different subtests of an IQ battery. Smaller differences are observed on subtests related to short-term memory while larger differences can be found on subtests of reasoning. This variation in mean Black/White differences between subtests cannot be explained by test bias, so there is a need to look for other causes. Charles Spearman (1923) suggested that each IQ battery subtest's *g* loading can be used to predict the difference in magnitude of the mean Black/White differences on the same subtest. Numerous studies have been carried out to test his so-called Spearman's hypothesis (Jensen, 1998) yielding an overwhelming amount of confirmations (Jensen, 1998).

Traditionally Spearman's hypothesis has been tested at the level of subtests of an IQ battery, but Rushton, in a series of papers, came up with a methodological innovation, namely he tested Spearman's hypothesis using item-level data from the Raven's Progressive Matrices. He argued that the *g*-loadedness of items could be measured by correlating the item score with the total score on the test. A recent paper by Gignac (2015) confirms that the Raven's has very high *g* loadings, indicating this should give a good measure of *g*. Rushton also argued that instead of the standardized differences between the scores of two groups on a subtest of an IQ battery, one could use the percentage of correct answers to each item, deduct the lower percentage. Then the remainder, together with the *g* loadings, can be correlated to test Spearman's hypothesis. Multiple empirical studies were carried out by Rushton and his co-authors, using participants from Africa and Serbia (Rushton, 2002; Rushton, Čvorović, & Bons, 2007; Rushton & Skuy, 2000; Rushton, Skuy, & Fridjhon, 2002; 2003). These studies generally showed modest support of Spearman's hypothesis with a mean unweighted correlation of approximately.30. The lower correlations may have been due to the sample sizes in Rushton's studies which generally were not large. te Nijenhuis and his co-authors recently carried out a series of studies generally using larger samples and also more samples. te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, and Repko (2015) tested Spearman's hypothesis using Libyan university students and adults. Groups from four different countries were made comparable on characteristics such as age and then compared to the Libyan group, this comparison yielding a mean-weighted *r* with a value of .73. te Nijenhuis et al. (2015) tested Spearman's hypothesis using Libyan secondary school children. The group from Libya and groups from seven different countries were made comparable on characteristics such as age. The other groups were then compared to the Libyan group and this yielded a mean-weighted *r* with a value of .61. te Nijenhuis, Grigoriev, and van den Hoek (2016) tested Spearman's hypotheses on diverse groups from Kazakhstan, namely Kazakh, Russian, Korean, Tatar, and Uzbek children. The different groups were compared to the Russian children yielding a mean-weighted *r* of .67. Lastly, te Nijenhuis et al. (2016) tested Spearman's hypothesis using Sudanese children and adolescents, which were compared to children and adolescents from ten different countries. The groups were made comparable on characteristics such as age and this yielded a sample-size-weighted *r* value of .70. So, the studies with larger samples by te Nijenhuis and co-authors generally led to higher correlations between *g* and *d*.

There is only one study in the literature not supporting Spearman's hypothesis (Diaz, et al.2012) compared a sample of Moroccans with a White Spanish sample. The authors describe how the Raven Standard Progressive Matrices was administered to 460 subjects (258 from Spain with a mean age of 25 and 202 from Morocco with a mean age of 26). The subjects consisted of university students studying a wide range of subjects and the staff of the university hall of residences, including administrative staff, cleaners, waiters, and cooks. When these samples had performed the test, the authors obtained further subjects through the social networks of these initial subjects (Díaz, et al.,

2012). The authors computed the differences in pass rates of the items with the *g* loadings of the items which yielded a correlation of *r* = -.20.

There have been multiple tests of Spearman's hypothesis at the item level with contradictory results, ranging from no support to strong support of the hypothesis (see also Wicherts (2017) for a critique). For a better understanding of the item level version of Spearman's hypothesis, we carried out an additional study which compares a large sample of children from Yemen with a Russian sample and a Kazakh sample. The test used from the study was the Raven's Progressive Matrices Plus. Our question is whether Spearman's hypothesis will hold up in Yemen. Will there be no support as found in Díaz's study, modest support as in Rushton's studies, or convincing support as in the studies by te Nijenhuis and his co-authors?

Method:

For this study we tested Spearman's hypothesis on children from Yemen comparing them to children from Kazakhstan and Russia. For this test we used the Method of Correlated Vectors at the item level, using the item scores from the Raven's Standard Progressive Matrices Plus (SPM+; Raven, 2008). We then combined the two scores using the meta-analytical software of Schmidt and Le (2014), computing a Harmonic-*N*-weighted mean correlation and the 80% credibility interval.

Instrument:

To test the hypothesis, the SPM+ was used. The SPM+ is an updated and improved version of the original Standard Progressive Matrices (SPM). The SPM+ is non-verbal test and has 60 different items which are divided over 5 sets with 12 items each. The items in the sets range from A to E; similar to the original SPM. The items become more difficult from item 1 to 12 and again from block A to E. The Raven's Progressive Matrices are commonly considered as an excellent test for measuring *g*. Therefore, we can expect the SPM+ to be a continuation of this tradition and should also be a high-quality test for testing Spearman's hypothesis.

Data:

For this test we used data from three different samples. Data on the Yemen sample was taken from Bakhiet, Al-Khadher, and Lynn (2015); this sample contained children ages 6-14 from public primary schools in the city of Dhamar and were representative for this area. For comparison purposes we only used children ages 7-13 (*N* = 1916) with a mean age of 9.78 since there were very few 14-year-olds.

Data on the Russian sample comes from unpublished data received from Grigoriev and Shibaev (2015); this sample contained children ages 6-17 with a mean age of 10, sampled from a single school in the city of Tomsk, Siberia. Since there was only a single 6-year-old, we compared this sample (*N* = 700) to the children ages 7-13 from the Yemen sample.

The sample from Kazakstan was taken from Griegoriev and Lynn (2014); the sample contained children between the ages of 8-18 from representative schools in the southern area of Kazakhstan. Due to the limited number of 8-year-olds we compared Kazakh children ages 9-13 (*N* = 426) with Yemen children ages 9-13 (*N* = 1405).

Calculating d:

The *d* or difference score for items is calculated by taking the score of the highest scoring group and deducting the score of the lower scoring group, the remainder is then divided by the standard deviation. The SPM+ doesn't use a mean score but a pass-rate, which is the percentage of correct answers of the entire group. As with other calculations of *d*, we take the percentage of the higher scoring group and deduct the percentage of the lower scoring group (Rushton, Skuy, and Fridjhon,

2003). As is common, we used the highest scoring group, the Russian group, as comparison group for all of the other groups.

Calculating g:

For the *g* loadings, we used the item-total correlation as a proxy for the degree to which the item measures general intelligence, we used the item-total correlation as a proxy (see Rushton, Skuy, and Fridjhon, 2003). Calculating the item-total correlation is done by taking the score for each subject and correlating it with their total score on the SPM+. Since the total score on the progressive matrices is a high-quality indicator of *g*, this measure is a good proxy for *g* loadings. The Russian data was used for these comparisons, since it was the largest white group, which is generally used to calculate *g* loadings. For these comparisons we used the *g* loadings calculated with the Russian data to test Spearman's hypothesis, since it was the largest White group, which is generally used to calculate *g* loadings. There were a few negative g loadings, since cognitive tasks cannot correlate negative with g, we changed these loadings to zero.

Results:

The outcomes from our testing of Spearman's hypothesis using item level data are reported in Table 1. Using the *g* loading from the Russian sample, the correlations between vectors range from .76 ($p < .0005$) to .86 ($p < .0005$) with a $N_{harmonic}$-weighted average correlation of .80 with a credibility interval ranging from .74 to .86.

Table 1:
Spearman's hypothesis tested using ethnic groups from Yemen and Kazakhstan

Comparison group	Yemen age (range)	Comparison group age (range)	r (d x g_{Yemen})	N_{Yemen}	$N_{comparison}$	$N_{harmonic}$
Kazakh	10.60 (9-13)	11.30 (9-13)	.76	1916	656	1955
Russian	10.03 (7-13)	9.78 (7-13)	.86	1405	426	1308

Note. $N_{harmonic}$ computed using the formula $\frac{4}{\frac{1}{n1}+\frac{1}{n2}}$ where N is the number of groups and where n1 and n2 are the amount of participants in group n1 and n2 respectively. The Russian g was used for all calculations. Both correlations p < .0005.

Discussion

We tested Spearman's hypothesis tested at the item level and expected the magnitude of the differences between groups on the level of items to be a function of the item's g loading, with larger differences on items with high g loadings and smaller differences on items with low g loadings. An empirical test by (Diaz, et al., 2012) comparing Spanish and Moroccans showed no support for Spearman's hypothesis, whereas other empirical studies showed modest to strong support. In the present study we tested whether another sample from the Arab world would replicate the outcomes of Díaz et al. of no support for Spearman's hypothesis.

In this study a large sample of children from Yemen was compared to a Russian sample and a Kazakh sample yielding outcomes of Spearman's hypothesis of, respectively, r = .86, and r = .76. Based on this finding and previous findings, we conclude that Spearman's hypothesis at the item level holds true. The only clear outlier in the literature remains the study by Díaz, et al. (2012) with an r = -.20.

A limitation of our study is that only samples of children and adolescents were compared to each other. Future studies should attempt to incorporate groups from different countries as well as more diverse age groups. This way the generalization of the findings can be tested. Furthermore, Spearman's hypothesis should be tested on a larger variety of tests besides the SPM and SPM+. Especially other non-verbal tests make good candidates for comparing groups from different countries, such as Raven's Coloured Progressive Matrices and Raven's Advanced Progressive Matrices.

References:

[1] Bakhiet, S., Al-Khadher, M., & Lynn, R . (2015). A study of means and sex differences on
[2] Raven's Standard Progressive Matrices Plus in Yemen. Mankind Quarterly, 55, 268-277.
[3] Díaz, A., Sellami, K., Infanzón, E., Lanzón, T., & Lynn, R. (2012). A comparative study of general intelligence in Spanish and Moroccan samples. The Spanish Journal of Psychology, 15, 526-532. http://dx.doi.org/10.5209/rev_SJOP.2012.v15.n2.38863
[4] Evers, A., te Nijenhuis, J., & van der Flier, H. (2005). Ethnic bias and fairness in
[5] personnel selection: Evidence and consequences. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), The Blackwell handbook of personnel selection (pp. 306-328). Oxford: Blackwell.
[6] Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. Intelligence, 52, 71-79. http://dx.doi.org/10.1016/j.intell.2015.07.006
[7] Grigoriev, A., & Lynn, R. (2014). A study of the intelligence of Kazakhs, Russians and
[8] Uzbeks in Kazakhstan. Intelligence, 46, 40-46. http://dx.doi.org/10.1016/j.intell.2014.05.004
[9] Grigoriev, A. & Shibaev, V. (2015) [Russian intelligence data]. Unpublished raw data.
[10] Jensen, A. R. (1998). The g factor: The science of mental ability. London: Praeger.
[11] Raven, J. (2008). Standard Progressive Matrices Plus Version and Mill Hill Vocabulary Scale
[12] Manual Manual. London: Pearson.
[13] Rushton, J. P. (2002). Jensen Effects and African/Colored/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. Personality and Individual Differences, 33, 65–70. http://dx.doi.org/10.1016/S0191-8869(02)00012-0
[14] 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. http://dx.doi.org/10.1016/j.intell.2006.09.002
[15] Rushton, J. P., & Skuy, M. (2000). Performance on Raven's Matrices by African and White university students in South Africa. Intelligence, 28, 251-265. http://dx.doi.org/10.1016/S0160-2896(00)00035-0
[16] Rushton, J. P., Skuy, M., & Fridjhon, P. (2002). Jensen effects among African, Indian,
[17] and White engineering students in South Africa on Raven's Standard Progressive Matrices. Intelligence, 30, 409-423. DOI: 10.1016/S0160-2896(02)00093-4
[18] Rushton, J. P., Skuy, M., & Fridjhon, P. (2003). Performance on Raven's Advanced
[19] Progressive Matrices by African, East Indian, and White engineering students in South Africa. Intelligence, 31, 123-137. http://dx.doi.org/10.1016/S0160-2896(03)00055-2
[20] Spearman, C. (1923). The nature of 'intelligence' and the principles of cognition**.** London: Macmillan.
[21] Schmidt, F. L., & Le, H. (2004). Software for the Hunter-Schmidt meta-analysis methods. Iowa City, IQ 42242: University of Iowa, Department of Management and Organization.
[22] te Nijenhuis, J., Al-Shahomee, A. A., van den Hoek, M., Allik, J., Grigoriev, A., & Dragt, J. (2015). Spearman's hypothesis tested comparing Libyan secondary school children with various other groups of secondary school children on the items of the Standard Progressive Matrices. Intelligence, 50, 118-124. http://dx.doi.org/10.1016/j.intell.2015.03.002
[23] te Nijenhuis, J., Al-Shahomee, A. A., van den Hoek, M., Grigoriev, A., & Repko, J. (2015). Spearman's hypothesis tested comparing Libyan adults with various other groups of adults on

the items of the Standard Progressive Matrices. Intelligence, 50, 114-117. http://dx.doi.org/10.1016/j.intell.2015.03.001

[24] te Nijenhuis, J., Bakhiet, S.F., van den Hoek, M., Repko, J., Allik, J., Žebec, M.S., (…), & Abduljabbar, A.S. (2016). Spearman's hypothesis tested comparing Sudanese children and adolescents with various other groups of children and adolescents on the items of the Standard Progressive Matrices. Intelligence, 56, 46-57. http://dx.doi.org/10.1016/j.intell.2016.02.010

[25] te Nijenhuis, J., de Jong, M. J., Evers, A ,van der Flier, H. (2004). Are cognitive differences between immigrant and majority groups diminishing? European Journal of Personality, 18, 405-434. http://dx.doi.org/10.1002/per.511

[26] te Nijenhuis, J., Evers, A., & Mur, J. P. (2000). The validity of the Differential Aptitude Test for the assessment of immigrant children. Educational Psychology, 20, 99–115. http://dx.doi.org/10.1080/014434100110416

[27] te Nijenhuis, J., Tolboom, E., Resing, W., & Bleichrodt, N. (2004). Does cultural background influence the intellectual performance of children from immigrant groups? The RAKIT Intelligence Test for immigrant children. European Journal of Psychological Assessment, 20, 10-26. http://dx.doi.org/10.1027/1015-5759.20.1.10

[28] te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. Journal of Applied Psychology, 82, 675–687. http://dx.doi.org/10.1037/0021-9010.82.5.675

[29] te Nijenhuis, J., & van der Flier, H. (2005). Immigrant-majority group differences on work-related measures: The case for cognitive complexity. Personality and Individual Differences, 38, 1213–1221. http://dx.doi.org/10.1016/j.paid.2004.08.004

[30] Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). Intelligence, 60, 26-38.

International Journal of Educational and Psychological Studies – Vol. 3, No. 3, 2018, pp. 720 - 726
e-ISSN 2520-4149 , p-ISSN 2520-4130
Available online at http:// www.refaad.com

Spearman's hypothesis tested in Yemen on the items of the Standard Progressive Matrices Plus: A reply to Díaz, Sellami, Infanzón, Lanzón, & Lynn (2012)

Salaheldin Farah Attallah Bakhiet[1,*], Jan te Nijenhuis[2,3], Michael van den Hoek[2]
Mohammed Mohammed Ateik Al-Khadher[1],and Shibaev Vladimir[4]

[1]King Saud University, Riyadh, Saudi Arabia
[2]Work and Organizational Psychology, University of Amsterdam, the Netherlands
[3]National Research Center for Dementia, Chosun University, Gwangju, Korea
[4]Far Eastern Federal University, Vladivostok, Russia
[*]Correspondence Author: Salaheldin Farrah Attallah Bakhiet, King Saud University, College of Education, Department of Special Education, KSA
E-mail: sLh9999@yahoo.com

الملخص:

تنص فرضية سبيرمان ببساطة على أن الاختلافات بين المجموعات في اختبار الذكاء هي دالة في الذكاء العام (g) على مستوى البنود، فإن هذا يعني أن حجم الاختلافات بين المجموعات أصغر عند البنود ذات التحميل المنخفض وأكبر على البنود ذات التحميل (g) االمرتفع. لم يظهر اختبار إجرائي من قبل Díaz , Sellami , Infanzón , Lanzón Lynn (2012) مقارنة بين الإسبان والمغاربة الذين طبق عليهم اختبار المصفوفات المتتابعة المعياري أي دعم لفرضية Spearman، في حين أظهرت دراسات تجريبية أخرى دعم إيجابي متواضع. في هذه الدراسة اختبرنا ما إذا كانت عينة أخرى من العالم العربي ستكرر نتائج Díaz et al. (2012) عدم دعم فرضية سبيرمان من خلال حساب العلاقة بين تحميل (g) والاختلافات في المجموعة على عناصر اختبارات الذكاء والتحقق مما إذا كانت تظهر علاقة سلبية. تمت مقارنة الأطفال من اليمن (العدد الكلي= ١٩١٦) مع مجموعة من الأطفال الروس (ن=٤٢٦) ومجموعة من الأطفال الكازاخستانيين (ن= ٦٥٦) على المصفوفات التقدمية القياسية زائد تنتج نتائج فرضية سبيرمان على التوالي، ر=٠,٨٦ (p <.0005) ، و ر٠,٧٦ (p <.0005) متوسط الارتباط المرجعي Nharmonic من٠,٨٠ لديه درجة صدق تتراوح من ٠,٧٤ إلى ٠,٨٦ استنادًا إلى النتائج، نستنتج أن فرضية Spearman تحمل حتى على مستوى البنود.

الكلمات المفتاحية: فرضية سبيرمان، اليمن، معدل الذكاء(g)، اختلافات المجموعة.