psych

MDPI

Article

Differential Item Functioning on Raven's SPM+ Amongst Two Convenience Samples of Yakuts and Russians

Vladimir Shibaev [1], Andrei Grigoriev [2], Ekaterina Valueva [2] and Anatoly Karlin [2,*]

[1] Vladivostok State University of Economics and Service, Vladivostok 690014, Russia; wladimirdw@gmail.com
[2] Institute of Psychology of Russian Academy of Sciences, Moscow 129366, Russia; andrey4002775@yandex.ru (A.G.); ekval@mail.ru (E.V.)
* Correspondence: alfatols@gmail.com

Received: 4 December 2019; Accepted: 3 January 2020; Published: 9 January 2020

Abstract: National IQ estimates are based on psychometric measurements carried out in a variety of cultural contexts and are often obtained from Raven's Progressive Matrices tests. In a series of studies, J. Philippe Rushton et al. have argued that these tests are not biased with respect to ethnicity or race. Critics claimed their methods were inappropriate and suggested differential item functioning (DIF) analysis as a more suitable alternative. In the present study, we conduct a DIF analysis on Raven's Standard Progressive Matrices Plus (SPM+) tests administered to convenience samples of Yakuts and ethnic Russians. The Yakuts scored lower than the Russians by 4.8 IQ points, a difference that can be attributed to the selectiveness of the Russian sample. Data from the Yakut (n = 518) and Russian (n = 956) samples were analyzed for DIF using logistic regression. Although items B9, B10, B11, B12, and C11 were identified as having uniform DIF, all of these DIF effects can be regarded as negligible (R2 <0.13). This is consistent with Rushton et al.'s arguments that the Raven's Progressive Matrices tests are ethnically unbiased.

Keywords: differential item functioning; Raven's SPM+; Russians; Yakuts

1. Introduction

National IQ estimates are made on the basis of psychometric measurements carried out in diverse cultural contexts [1–4]. These measurements are often obtained from Raven's Progressive Matrices tests. In Richard Lynn's book, *Race Differences in Intelligence: An Evolutionary Analysis* (2006), some 38 out of the 73 measurements of indigenous European IQ (Table 3.1 therein) and 32 out of the 57 measurements of Sub-Saharan African IQ in Africa (Table 4.1 therein) were made using Raven's Matrices. Seven of the 10 articles on new national IQ measurements published in the journal Mankind Quarterly in 2018 were done using Raven's Matrices (in one case, in combination with the WAIS-R). It is, therefore, unsurprising that Raven's Matrices have often figured in studies on measurement invariance (discussions of test bias) between respondents from different ethnic/national groups.

Many of these studies were authored by J. Philippe Rushton [5–8], and typically used two methods to assess test bias. First, they calculated the product-moment and rank-order correlations between the pass rates of the test items in the ethnic samples. These were usually equal to or higher than 0.9. Second, they used the method of correlated vectors, in which one vector was the difference in pass rates between the two groups, while the other vector was the vector of the item–total correlations. These correlations were typically significant. Summarizing the results of these studies—as well as that of Owen (1992) [9]—Rushton and Arthur Jensen concluded that they "found almost identical item structures in Africans, Whites, and East Indians on the Progressive Matrices. Items found difficult by

one group were difficult for the others; items found easy by one group were easy for the other. The item–total score correlations for Africans, Whites, and East Indians were also similar, indicating that the items measured similar psychometric constructs in all three groups" [10].

However, these approaches have been brought into question [11]. The critics noted, in particular, that the method of correlated vectors "does not address the issue of measurement invariance" [12]. It is worth noting that the method of correlated vectors was criticized when applied to testing Spearman's hypothesis [13,14], and in studies aiming to show the heritability of group differences in intelligence [15].

Measurement invariance is better assessed by testing for differential item functioning (DIF). DIF appears when subjects with the same ability level have different probabilities of answering a particular item correctly. An item can be considered as DIF if either the difficulty level (uniform DIF) or discriminative power (nonuniform DIF) is different in the two groups (for details see Facon at al., 2011 [16]). DIF can be examined using confirmatory factor analysis [17], item response theory (IRT) methods [18], or non-IRT methods such as the Mantel–Haenszel method or logistic regression [19].

To date, there have been few studies in which DIF analysis was applied to interethnic comparisons using Raven's Matrices. Wicherts et al. in 2010 [12] noted that they were aware of only one such study by Taylor (2008), in which item response theory (IRT) was used in the analysis [20]. We also managed to find another pre-2010 study by Vanderpool and Catano (2008), which involved a DIF analysis of Raven's Matrices (and other tests) using logistic regression [21].

In Taylor's study, there were 4 items in the Raven's Standard Progressive Matrices (SPM) and 5 items in the Raven's Advanced Progressive Matrices (APM) that showed the presence of DIF when comparing blacks and whites. In the Vanderpool and Catano study, 2 items in Raven's SPM displayed DIF in a comparison of Native Americans and (mostly) whites. The authors of both studies believe that the Raven's SPM test was unbiased against their respective focus groups.

These studies cannot be considered definitive for several reasons. First, none of them corrected for multiple comparisons; if that had been done, especially in studies using logistic regression for DIF detection [22], it is entirely possible that DIF would have disappeared for some or all of the items. Second, sample sizes were modest: In the first study, there were 200 blacks and 178 whites taking the SPM, and 67 blacks and 115 whites taking the APM; in the second study, there were 101 Native Americans and 108 whites. This implies a sufficiently high likelihood of a second mistake: Some items that are characterized by DIF may have remained undetected.

Consequently, there is still much work to be done assessing Rushton et al.'s conclusions regarding the unbiasedness of Raven's Matrices tests applied to interethnic comparisons through DIF analysis. This problem has become all the more germane in light of DIF analyses of Raven's Matrices tests conducted since 2010 applied to other group differences, such as age, sex, and intellectual disabilities [16,23–25]. Some items with minor DIF were found in these studies. For example, in the Facon et al. (2011) study, children and adolescents with an intellectual disability were compared with typical children with the Raven's Colored Progressive Matrices test [16]. Using logistic regression, the authors identified 12 items out of 36 as functioning differentially between the two groups, out of which 10 items exhibited negligible DIF, and only 2 items exhibited moderate DIF.

In this paper, we present a DIF analysis of a Raven's SPM+ taken by a sample of Yakuts (students in the Sakha Republic) and ethnic Russians (pupils at a Tomsk school and children from a Vladivostok summer camp).

2. Materials and Methods

Materials for this study were gathered in the Sakha Republic (Yakutia), Tomsk, and a Vladivostok summer camp.

Data on respondents who identified as Yakuts were gathered in the Sakha Republic (Yakutia) from the following educational institutions: North-Eastern Federal University (NEFU), the Republican

Lyceum, the Sakha Polytechnic Lyceum, School No.31 in Yakutsk, as well as from the Vilyuysk Gymnasium and School No.3 in Vilyuysk.

Data on ethnic Russians were gathered from Tomsk and Vladivostok. In Tomsk, the study was carried out at the G.A. Psakhye Academic Lyceum where 957 pupils were tested. Most of the 55 children tested at the Vladivostok summer camp were evacuees from the Amur river floods that year and maintained their permanent residence in villages close to Khabarovsk (most of them came from Bichevaya village).

The overall sample size was composed of 1531 respondents, of whom 519 were Yakuts and 1012 were ethnic Russians, 764 were male and 767 were female. Ages ranged from 6 to 29 years, with an average of 12.16 years and an SD of 2.94 years.

Raven's SPM+ tests were administered in a paper-and-pencil format. There were no time limits, but the test period usually lasted for the duration of a school lesson (45 min).

Both samples were conventional and non-representative to regional populations, although the Yakut one, perhaps, to a lesser extent since data were drawn from the Sakha Republic's capital city (Yakutsk) and from a small town of 11,000 people (Vilyuysk), as well as from a variety of educational institutions. For instance, in Vilyuysk, data were collected from both its best and worst school, as assessed by the local education department. The Russian sample, on the other hand, should be considered selective, since the overwhelming bulk of it was drawn from just one Tomsk school, the G.A. Psakhye Academic Lyceum. Tomsk is the capital of an oblast that occupies the 5th place amongst 85 Russian regions according to results from Internet testing of cognitive abilities [26], and lyceum students were, on average, more cognitively able than students at ordinary schools. The subsample from Vladivostok was probably more representative of the regional population, but it constituted only a small portion of the Russian sample.

One must, therefore, be cautious about generalizing from the results of this study.

3. Results

In the Yakut sample, one 19-year-old girl correctly answered only 4 items. This was well below what she was expected to have gotten answering randomly, thus her answers were excluded from the analysis. The Tomsk sample had many instances of probable cheating, as deduced from the presence of identical answer sheets, often in sequential order from the same class groups. There were 28 pairs of such identical answer sheets, which we excluded in their entirety (i.e., 56 answer sheets). Consequently, the final sample consisted of 1474 respondents, of whom 518 were Yakuts and 956 were Russians.

The number of respondents in each age group in the samples of Yakuts and Russians, and their average scores and SDs, are given in Table 1.

The weighted mean d was calculated. The weights were the harmonic means of two sample sizes. It was equal to −0.32, which corresponded to ~ 4.8 IQ points (on the standard IQ scale with a mean of 100 and an SD of 15).

Two other studies provided data from which the difference between the Yakutia and Tomsk region could be estimated. The first of these [26] were based on results from an anonymous Internet test that aimed to assess eligibility for contract military service. According to that study, Tomsk oblast got an average score of 20.897 (SD = 5.892, n = 1442), while Yakutia got 19.793 (SD = 5.762, n = 647). The resulting effect size was 0.187, or ~ 2.8 IQ points (on the standard IQ scale with a mean of 100 and an SD of 15). Another study was the study by Lynn, Cheng, and Grigoriev (2017) that analyzed test results in the Russian regions covered by the Programme for International Student Assessment (PISA) 2015 [27]. In this study, Tomsk got an average score of 480.6 (SD = 72.6, n = 79), while Yakutia got 469.1 (SD = 65.0, n = 96). This translated to a difference of ~2.5 IQ points or d = 0.168. However, because our results were obtained from convenience samples, they could not be compared with these other estimates.

Table 1. The number of respondents, means, and SD's of test scores, and Cohen's d across age groups for Russians and Yakuts.

Age	Yakuts			Russians			D [1]	P
	N	Mean	SD	N	Mean	SD		
<7	-	-	-	76	26.9	6.2	-	-
8	2	27.5	2.1	105	28.2	4.9	−0.15	0.713
9	34	25.7	6	115	29.7	6	−0.66	0.001
10	21	29.1	4.5	125	32.1	4.7	−0.64	0.01
11	30	32.9	4	121	32.2	4.5	0.16	0.378
12	82	32.4	5.9	89	34.8	4	−0.48	0.002
13	62	38	4.5	82	36.6	5.4	0.28	0.091
14	59	36.7	4.2	96	39.9	5.1	−0.67	0
15	79	38.1	5.4	77	39.9	4.9	−0.35	0.029
16	61	40.2	5.6	59	42	5.6	−0.31	0.091
17	56	42.1	7	11	42.5	6	−0.07	0.813
18+	32	39.5	5.2	-	-	-	-	-

[1] d was calculated using the formula: $(m_y - m_r)/sqr(((n_y - 1) \times s_y^2 + (n_r - 1) \times s_r^2) / (n_y + n_r - 2))$ where m_y and m_r are the means for the Yakut and Russian samples, s_y and s_r are the standard deviations for the Yakut and Russian samples, n_y and n_r are the numbers of respondents in the Yakut and Russian samples, thus, negative d represents a Russian advantage.

4. DIF Analysis

Ten items (A1–A6, A9, B1–B3) were excluded from the analysis because they showed a kurtosis value of more than 30, and more than 97% of responses to each of them were correct.

Data processing was done in R using the package difR [19]. Both uniform and nonuniform DIF were evaluated using logistic regression. Procedures included item purification and Holm corrections for multiple comparisons. The purpose of item purification was to avoid bias when identifying DIF by the inclusion of DIF items in the anchor items set (in the case of logistic regression, this was the total test score used as a proxy for ability level). It was an iterative process, in which all items that exhibited DIF on the first step were excluded from the anchor items and were sequentially tested for DIF. This process was repeated until two consecutive steps gave identical results, or a set number of iterations was exceeded. In Table 2, we list test items for which DIF was identified, as well as the Holm adjusted p-value of the likelihood ratio test, the type of DIF effect, and the value of Nagelkerke's ΔR^2 as the effect size. According to Zumbo and Thomas (1997), ΔR^2 of less than 0.13 was considered as negligible, of more than 0.26 as large, and between those two values as moderate [28]. Jodoin and Gierl (2001) suggested using 0.035 and 0.070 as cutoff values [29].

Table 2. Items flagged as having differential item functioning (DIF).

Item	Adjusted *P*-Value	DIF Type	Effect Size
B9	0.0068	uniform	0.02
B10	0.029	uniform	0.01
B11	0	uniform	0.03
B12	0	uniform	0.01
C11	0	uniform	0.03

Five items demonstrated uniform DIF, and no items demonstrated nonuniform DIF. Even according to Jodoin and Gierl's more liberal criteria, all DIF effects in our study—despite their significance—can be regarded as negligible (ΔR^2 <0.035). On all of these items, the Russian sample gave a higher number of correct answers.

Figure 1 presents the logistic curves for DIF items.

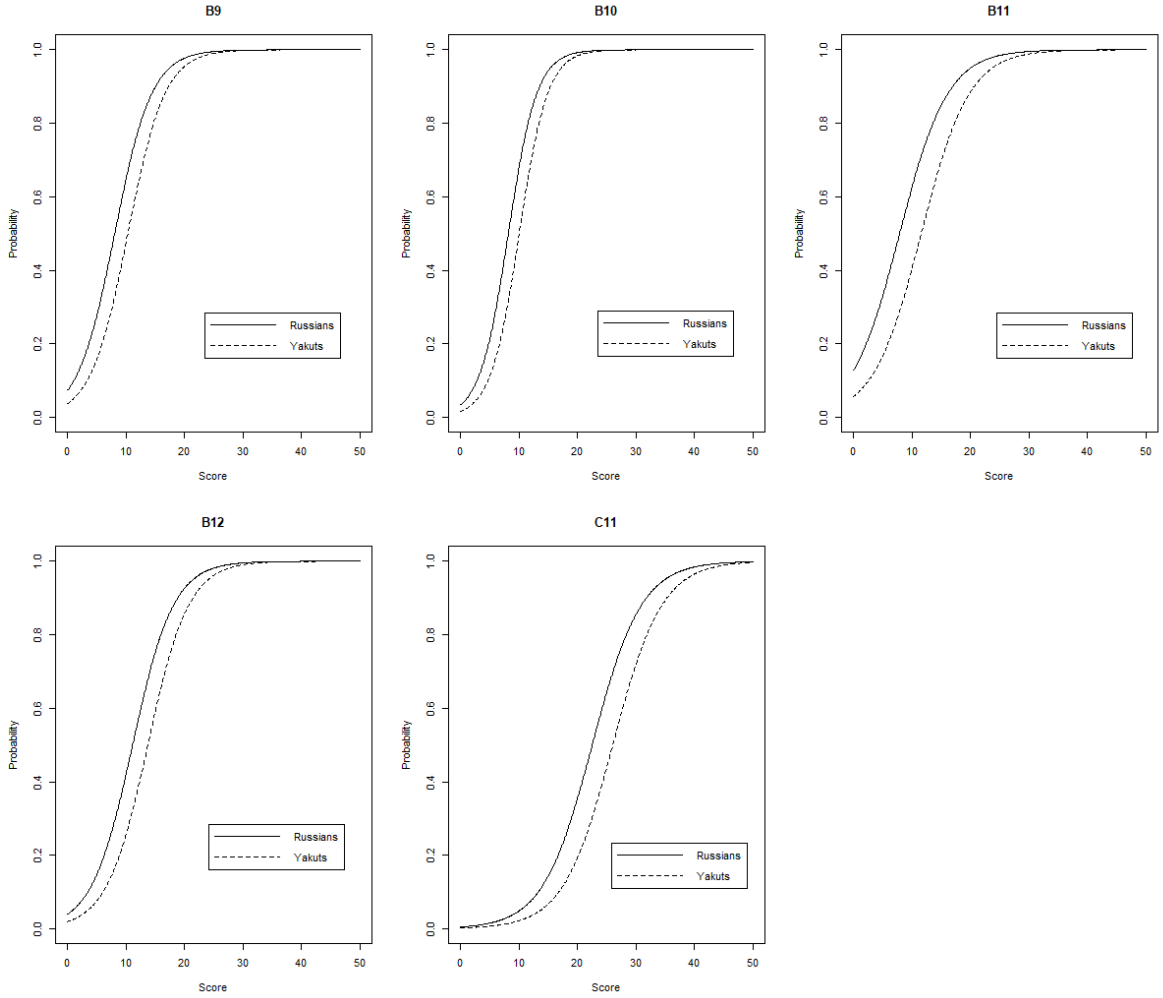

Figure 1. Logistic curves for DIF items.

The test data for Yakuts and Russians is repeated below in Table 3, but this time without items that displayed DIF (B9–B12, C11).

Table 3. The number of respondents, means, and SD's of test scores, and Cohen's d across age groups for Russians and Yakuts.

Age	Yakuts			Russians			D [1]	P
	N	Mean	SD	N	Mean	SD		
<7	-	-	-	76	24.2	5.4	-	-
8	2	24.5	2.1	105	25.2	4.2	−0.17	0.712
9	34	23.6	4.7	115	26.6	5.1	−0.6	0.002
10	21	26.6	3.5	125	28.7	4.1	−0.52	0.021
11	30	29.7	3.7	121	28.9	4.1	0.2	0.302
12	82	29.4	4.9	89	30.8	3.7	−0.33	0.035
13	62	34.2	4.1	82	32.7	4.7	0.32	0.051
14	59	33.1	4	96	35.7	4.8	−0.57	0
15	79	34.5	4.9	77	35.8	4.7	−0.27	0.094
16	61	36.3	5.1	59	37.7	5.3	−0.26	0.16
17	56	38.1	6.5	11	38.1	5.6	0	0.992
18+	32	36.1	5.1	-	-	-	-	-

[1] d was calculated using the formula: $(m_y - m_r)/sqr(((n_y - 1) \times s_y^2 + (n_r - 1) \times s_r^2)/(n_y + n_r - 2))$ where m_y and m_r are the means for the Yakut and Russian samples, s_y and s_r are the standard deviations for the Yakut and Russian samples, n_y and n_r are the numbers of respondents in the Yakut and Russian samples, thus, negative d represents a Russian advantage.

The weighted mean d was calculated. The weights were the harmonic means of two sample sizes. It was equal to −0.24, which corresponded to ~ 3.6 IQ points.

Data and code are available in the Supplementary Materials.

5. Discussion

Our DIF analysis on a Raven's SPM+ test administered to convenience samples of Russians and Yakuts identified 5 items out of 50 (10%) with significant, though small, DIF. Four of these items (B9–B12) represent similar tasks. Van der Ven and Ellis (2000) characterize the items B8–B12 in SPM, which closely resemble the corresponding items in SPM+, as demanding analogical reasoning, unlike items B1–B6 of items of set A, which must be solved according to some Gestalt continuation rule [30]. Do they load on any specific ability that might be more strongly expressed in one of the studied ethnic groups? Perhaps subsequent studies can answer this question.

However, our results sooner support the idea that comparisons between diverse groups show minimal bias when Raven's SPM+ is used. Although this study used convenience samples, this should not be an issue, since DIF is more likely, ceteris paribus, to appear in comparisons between more heterogeneous groups. Refraining from an assessment of the methods used in Rushton and Jensen [10], we can say that our results are generally consistent with their conclusions that the Raven's Matrices preserve their construct validity when applied to diverse ethnic groups.

Given the insufficient representativeness of our samples, we cannot claim that the Russian advantage in IQ scores relative to Yakuts that we observed is representative of the differences between the two populations at large, especially considering that a comparison of more similar samples showed no significant differences between Russians and Yakuts [31]. More likely, the difference obtained in the present study, or at least a large part of it, is due to the selectiveness of the Russian sample. It might also be that the revealed minimal bias is due, to some extent, to educational selectivity. Further investigations are needed.

Supplementary Materials: The following are available online at http://www.mdpi.com/2624-8611/2/1/5/s1, Title: Data and code.

Author Contributions: Data curation, E.V.; Funding acquisition, V.S.; Investigation, V.S.; Methodology, A.G.; Resources, V.S.; Supervision, A.G.; Visualization, E.V.; Writing—original draft, A.G. and A.K.; Writing—review & editing, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: The fieldwork component of this research received support from the Ulster Institute for Social Research. The study was conducted according to the state assignment of the Ministry of Science and Higher Education of the Russian Federation No 0159-2019-0010.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lynn, R.; Vanhanen, T. *IQ and the Wealth of Nations*; Praeger Publishers/Greenwood Publishing Group: Westport, CT, USA, 2002; p. 298. ISBN 9780275975104.
2. Lynn, R.; Vanhanen, T. *IQ and Global Inequality*, 1st ed.; Washington Summit Publishers: Augusta, GA, USA, 2006; ISBN 9781593680244.
3. Lynn, R.; Vanhanen, T. *Intelligence: A Unifying Construct for the Social Sciences*; Ulster Institute for Social Research: London, UK, 2012; ISBN 9780956881182.
4. Lynn, R. *Race Differences in Intelligence: An Evolutionary Analysis*; Washington Summit Books: Augusta, GA, USA, 2006; ISBN 9781593680206.
5. Rushton, J.P.; Skuy, M. Performance on Raven's Matrices by African and White University Students in South Africa. *Intelligence* **2000**, *28*, 251–265. [CrossRef]
6. Rushton, J.P. Jensen Effects and African/Coloured/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. *Pers. Individ. Differ.* **2002**, *33*, 1279–1284. [CrossRef]
7. Rushton, J.P.; Skuy, M.; Fridjhon, P. Jensen Effects among African, Indian, and White engineering students in South Africa on Raven's Standard Progressive Matrices. *Intelligence* **2002**, *30*, 409–423. [CrossRef]

8. Rushton, J.P.; Skuy, M.; Bons, T.A. Construct validity of Raven's advanced progressive matrices for African and non-African engineering students in South Africa. *Int. J. Sel. Assess.* **2004**, *12*, 220–229. [CrossRef]
9. Owen, K. The suitability of Raven's standard progressive matrices for various groups in South Africa. *Pers. Individ. Differ.* **1992**, *13*, 149–159. [CrossRef]
10. Rushton, J.P.; Jensen, A.R. Thirty years of research on race differences in cognitive ability. *Psychol. Public Policy Law* **2005**, *11*, 235–294. [CrossRef]
11. Wicherts, J.M. Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence* **2017**, *60*, 26–38. [CrossRef]
12. Wicherts, J.M.; Dolan, C.V.; Carlson, J.S.; van der Maas, H.L.J. Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn Effect. *Learn. Individ. Differ.* **2010**, *20*, 135–151. [CrossRef]
13. Raven, J. Testing the Spearman-Jensen Hypothesis Using the Items of the RPM. Available online: http://eyeonsociety.co.uk/resources/testingSJHyp.pdf. (accessed on 13 December 2019).
14. Wicherts, J.M. Ignoring psychometric problems in the study of group differences in cognitive test performance. *J. Biosoc. Sci.* **2018**, *50*, 868–869. [CrossRef]
15. Wicherts, J.M.; Johnson, W. Group differences in the heritability of items and test scores. *Proc. Biol. Sci.* **2009**, *276*, 2675–2683. [CrossRef] [PubMed]
16. Facon, B.; Magis, D.; Nuchadee, M.-L.; De Boeck, P. Do Raven's Colored Progressive Matrices function in the same way in typical and clinical populations? Insights from the intellectual disability field. *Intelligence* **2011**, *39*, 281–291. [CrossRef]
17. Nye, C.D.; Drasgow, F. Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *J. Appl. Psychol.* **2011**, *96*, 966–980. [CrossRef] [PubMed]
18. Meade, A.W. A taxonomy of effect size measures for the differential functioning of items and scales. *J. Appl. Psychol.* **2010**, *95*, 728–743. [CrossRef]
19. Magis, D.; Béland, S.; Tuerlinckx, F.; De Boeck, P. A general framework and an R package for the detection of dichotomous differential item functioning. *Behav. Res. Methods* **2010**, *42*, 847–862. [CrossRef]
20. Taylor, N. Raven's Standard and Advanced Progressive Matrices among adults in South-Africa. In *Uses and Abused of Intelligence. Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics*; Raven, J., Raven, J.C., Eds.; Royal Fireworks Press: Unionville, NY, USA, 2008; pp. 371–391.
21. Vanderpool, M.; Catano, V.M. Comparing the Performance of Native North Americans and Predominantly White Military Recruits on Verbal and Nonverbal Measures of Cognitive Ability. *Int. J. Sel. Assess.* **2008**, *16*, 239–248. [CrossRef]
22. Kim, J.; Oshima, T.C. Effect of Multiple Testing Adjustment in Differential Item Functioning Detection. *Educ. Psychol. Meas.* **2013**, *73*, 458–470. [CrossRef]
23. Chiesi, F.; Ciancaleoni, M.; Galli, S.; Primi, C. Using the Advanced Progressive Matrices (Set I) to assess fluid ability in a short time frame: An item response theory-based analysis. *Psychol. Assess.* **2012**, *24*, 892–900. [CrossRef]
24. Chiesi, F.; Ciancaleoni, M.; Galli, S.; Morsanyi, K.; Primi, C. Item Response Theory analysis and Differential Item Functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learn. Individ. Differ.* **2012**, *22*, 390–396. [CrossRef]
25. Waschl, N.A.; Nettelbeck, T.; Jackson, S.A.; Burns, N.R. Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Pers. Individ. Differ.* **2016**, *100*, 157–166. [CrossRef]
26. Sugonyaev, K.; Grigoriev, A.; Lynn, R. A New Study of Differences in Intelligence in the Provinces and Regions of the Russian Federation and Their Demographic and Geographical Correlates. *Mank. Q.* **2018**, *59*, 31–37.
27. Lynn, R.; Cheng, H.; Grigoriev, A. Differences in the Intelligence of 15 Year Olds in 42 Provinces and Cities of the Russian Federation and Their Economic, Social and Geographical Correlates. *Mank. Q.* **2017**, *57*, 659–668.
28. Zumbo, B.D.; Thomas, D.R. *A Measure of Effect Size for A Model-Based Approach for Studying DIF*; University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science: Prince George, BC, Canada, 1997.
29. Jodoin, M.G.; Gierl, M.J. Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection. *Appl. Meas. Educ.* **2001**, *14*, 329–349. [CrossRef]

30. Van der Ven, A.H.G.S.; Ellis, J.L. A Rasch analysis of Raven's standard progressive matrices. *Pers. Individ. Differ.* **2000**, *29*, 45–64. [CrossRef]
31. Shibaev, V.; Lynn, R. The Intelligence of Yakuts and Ethnic Russians in Yakutia. *Psychol. J. High. Sch. Econ.* **2016**, *13*, 678–682.